

Theorem: Single-link clustering method is order-independent.

Preliminary: Assume we have a list of pairs (p_1, p_2, \dots, p_k) sorted at non-decreasing order according to their similarity values. Let $n+1$ of them have the same similarity value s (p_t to p_{t+n} for some t and n , $t+n \leq k$).

Initially, -if the number of documents is m - we have m clusters each having only one document.

Lemma: Swapping any two adjacent pairs does not change the structure of the dendrogram if single-link clustering method is used.

Proof of Lemma: Let these two pairs are p_i and p_{i+1} . We have two lists now, first one is the original list and the other one is formed by swapping i^{th} and $(i+1)^{\text{th}}$ pairs of the original list.

Since the pairs are same up to p_{i-1} , the structures of the intermediate clusters are same before processing i^{th} pairs. If we can show that the structures of the dendrograms are same after processing $(i+1)^{\text{th}}$ pairs, than it is obvious that the structures of the dendrograms after processing the whole list will be the same, because the rest of the lists are identical.

According to the clusters just before i^{th} pair is processed, let the name of the clusters that the documents of p_i belongs to are A and B. (e.g. p_i is d_4 - d_7 with similarity value 0.6, than the intermediate cluster that contains d_4 is A, and the intermediate cluster that contains d_7 is B.) Similarly, C and D are the intermediate clusters that contain the documents of p_{i+1} . (Obviously, some of them may refer to the same cluster.)

According to the first list, we first merge A&B than C&D. According to the second list, we merge C&D and than A&B.

- If A and B are same, than we do not perform any merge for p_i , because they are already merged. So, processing p_i or p_{i+1} first does not change the structure obviously. Similarly, equality of C and D does not change the structure.
- If one of A&B refers to the same cluster with one of C&D ($A=C$ or $A=D$ or $B=C$ or $B=D$), again we will have the same dendrograms after $(i+1)^{\text{th}}$ pair, because both order of processing will yield merging three clusters with similarity value s . (e.g. $A=C$, according to the first list, we merge A&B with value s than merge D to A&B again with value s . According to the second list, we merge A&D and than merge B to A&D with value s . In both order, A&B&D are merged to each other with same similarity value s .)
- If $A=C$ and $B=D$ (or $A=D$ and $B=C$), again it is obvious that the order of them does not change the structure. Because both of them cause the same merge operation. So $(i+1)^{\text{th}}$ pair will be dummy since they are already merged.
- If there is no equality (A, B, C and D are different), than they offer independent merge operations. The structure will not be affected.

As we have seen, the structure does not change in all cases. Proof of lemma is over.

Proof of Theorem: We will prove the theorem by induction on the number of pairs that have the same similarity value.

If reordering a sub-list of length n that have the same similarity value s does not change the structure of the dendrogram, than reordering a sub-list of length $n+1$ also does not change the structure.

The base case is for $n=1$ (There are two pairs: p_t and p_{t+1}). It is shown in lemma. Reordering a portion of length 2 (swapping) does not change the structure of the dendrogram.

It is provided (by the assumption) that reordering the pairs starting from p_{t+1} up to p_{t+n} does not change the structure. For all $n!$ arrangements, the structure is same.

By swapping adjacent pairs, we can move any pair to the beginning (t^{th} position), rearrange the rest n pair and this *will not change the structure of the dendogram*. In this way, we can obtain all $(n+1)!$ arrangements by moving each p_i ($t \leq i \leq t+n$) to the beginning and rearranging the rest n pair. (We have $(n+1)$ different pairs to move to the beginning, and for each we have $n!$ arrangements. Totally, we can obtain $(n+1) \cdot n! = (n+1)!$ arrangements.)

Here is an example. Assume we have 3 pairs with same similarity value s ($n=2$). Let them be p_t , p_{t+1} and p_{t+2} .

$$[p_t, p_{t+1}, p_{t+2}] \equiv [p_t, p_{t+2}, p_{t+1}] \equiv [p_{t+2}, p_t, p_{t+1}] \equiv [p_{t+2}, p_{t+1}, p_t] \equiv [p_{t+1}, p_{t+2}, p_t] \equiv [p_{t+1}, p_t, p_{t+2}]$$

In each step above, we reordered a portion of length 2 (swapped adjacent pairs). At last, we obtained all $3!$ arrangements. Similarly, we can obtain $4!$ arrangements using this, than we can obtain $5!$ arrangements, and so forth...

Ramazan YILMAZ