Normalization in KMeD2 Temporal Indexing Module

Sanghyun Park

December 27, 2000

Similarity search in sequence databases can be performed either on the *raw sequences* or on the *normalized sequences.* In single dimensional sequences, normalization is required to support the shape preserved transformations such as shifting and scaling. In multi dimensional sequences, normalization makes all dimensions have the same data distribution; thus makes every dimension have the same amount of contribution to determine the similarity.

Let us consider the normalization in the database of m single dimensional data sequences, $\vec{X}_1, ..., \vec{X}_m$. n_i denotes the number of elements in $\vec{X}_i$ and $\vec{X}_{ij}$ the j-th element of $\vec{X}_i$. We first compute the average Avg and the standard deviation $Stddev$ from the set of data sequences. Then, each element $\vec{X}_{ij}$ is converted to $\frac{\vec{X}_{ij}-Avg}{Stddev}$. After normalization, the set of data sequences has the normal distribution with the average value 0 and the standard deviation 1.

$$Avg = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\vec{X}_{ij}}{\sum_{i=1}^{m}|n_i|}$$

$$Stddev = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\vec{X}_{ij} - Avg)^2}{\sum_{i=1}^{m}|n_i|}}$$

The normalization process for multi-dimensional sequences is not much different from the one for single-dimensional sequences. We derive the set of single-dimensional sequences for each dimension and apply the normalization process described above. After normalization, every dimension has the normal distribution with the average value 0 and the standard deviation

1. Therefore, all dimensions have the same contribution to determine the similarity. Once we achieve the property of the same contribution, we can assign each dimension the appropriate weight factor corresponding to its relative importance.