

Using Self-Organizing Maps to Extract Semantic Information from a Document Set

The Self-Organizing Map is a simple form of neural computing originally developed by Teuvo Kohonen. The map uses a grid or matrix of nodes each containing a vector or array of values between 0 and 1. Each item in the vector corresponds to a variable or parameter of the information space being studied.

As an example, let's say we are looking at keywords for a set of documents. For keyword extraction from a set of documents, first we remove "stop words" – those common articles like "the," "an" and "her" that carry little unique information. Next extract from the total list of words the words most likely to represent a given document.

Counter to conventional wisdom, word frequency alone is not a good predictor for picking a document out of a set. Instead we look for relatively unique keywords that would pick a particular document out of the set. On the other hand, the keyword needs to appear in at least one other document. After identifying the full set of keywords likely to find individual documents, we can rank these and pick the top terms, say the top 50 terms. This list defines the vector parameters or variables, in this case keyword terms. For each document a vector is defined with either a 1 or 0 for each item depending on whether the term appears in that document.

Similarly, each node in the map's grid is given a vector. The values for each node/vector/term are initially set to randomly generated, very small values, and randomly assigned to nodes in the grid. The set of document vectors can then be mapped to the grid, each document to the nearest node based on vector similarity.

Randomly pick a document vector and map it. For each map node/vector/item take the average of the map vector value and the document vector value, apply some weighting and set the map's vector to the new values. Also apply weighted modifications to map vectors in the surrounding area. One way to do this is to weight the modification based on a Gaussian function.

Now map the rest of the document vectors. Repeat this iteratively, progressively shrinking the size of the neighborhood and the weights applied.

When done, each document vector will again map to a particular map vector, but the map vector will contain values for semantically related keywords, garnered from the set and process that may not have appeared in the original document. In addition closely related documents will cluster together geographically on the map.

Using the Self-Organizing Map, we can define a set of related terms and relational similarity at a number of levels. We get a network of term similarity across the document set, an ability to define document clusters, the ability to identify documents from within the set that are semantically central or peripheral and the ability to define document similarity as a whole, or along any individual term parameter.