Granular Computing and Web Processing: Representing Documents in Polyhedron

T.Y. Lin
San Jose State University

Granular computing (GrC) is a pragmatic approach to Data and Knowledge engineering. Roughly, it is a methodology that involves elements (data) and subsets (knowledge). We will illustrate the idea via a web application.

In traditional web processing, we often represent a document by a set of keywords. In GrC, let us call it granular representation, we include the knowledge. More precisely, a document is not only represented by its keywords, but also by its granules. Here by a granule we mean a keyword association, which is a set of frequent co-occurring near by keywords. For a trivial example, the association, “Wall street,” as a financial concept, is in the granular representation.

The granular representation has an interesting geometric interpretation. We can regard the set of keywords as a set of vertices, and the set of keyword associations as a set of simplexes. Interestingly in such a translation, the a priori principle is converted into the closed condition of simplical complexes. In other words, the collection of keywords and keyword associations is an abstract simplicial complex in algebraic topology. Note that a simplicial complex is a combinatorial structure of polyhedron. So in this fashion the granular representation represents a document by a polyhedron. Based on such a polyhedron, documents can be clustered into various categories.

Tsau Young (T.Y.) Lin received his Ph.D. from Yale University, and now is a Professor in the Department of Computer Science, San Jose State University. Currently, he is also the Chapter Chair of IEEE Computer Society in Silicon Valley. He has served as editor, associate editor, member of the advisory/editorial board of several international journals, and chair, co-chair and member of program committees of conferences. His interests include approximation retrievals, data mining, data warehouse, data security, and computing methodology (granular computing, rough set, and soft computing).