Formal Concept Analysis and Document Clustering via Granular Computing

蔣以仁
Tsau Young Lin;I-Jen Chiang

Abstract

A text/web document is a knowledge representation of a human idea (a structured set of thoughts). This paper refines TFIDF and extended TFIDF(ETFIDF)[16]; These values really measures the co-occurrences of tokens. The ETFID captures the semantic more accurately. Tokens with high TFIDF values are called keywords. The sets of (n+1) Co-occurring keywords with High ETFIDF are called n-granules. The collection of keywords and n-granules can be interpreted geometrically; they form a non-closed simplicial complex. The corresponding non-closed polyhedron is called latent semantic space(LSS). LSS is a geometric knowledge base that provides the semantic to search engine.