Keyword Field Cleaning Through ClusterSuite: A Term-Clumping Tool for VantagePoint Software

Alan Porter, Stephen Carley, and John O'Brien

Georgia Institute of Technology, Atlanta, Georgia
E mail: alan.porter@isye.gatech.edu

Keywords: data cleaning, term clumping, text mining, VantagePoint software applications

Abstract
In a general sense, term clumping macros perform dimension reduction on a list, making it more approachable and enabling the user to see the forest from the trees. They minimize noise and maximize prominent topics, which enables the user to more quickly extract meaning from large amounts of text. More specifically, term clumping macros indicate (i) how closely related two or more terms are, (ii) a good name for a group that includes common terms, and (iii) the nature of the relationship among terms (e.g. parent-child, siblings, etc).

Currently, multiple methods exist for dimension and noise reduction on a list within the VantagePoint software package. Automated processes exist in formats as varied as thesauri files, VBA scripts, standalone programs, and more. ClusterSuite is an application written in VBA programming language with an HTML user-interface that runs as a script within VantagePoint. More specifically, ClusterSuite condenses, streamlines, and sequentially executes previously existing VantagePoint thesauri and scripts. At present, ClusterSuite organizes its parts into three phases. Phase I is the currently most developed phase. It executes five thesauri and one list cleaning macro. Phase II runs one of two term-clumping macros. Phase III is designed to eliminate extreme list components based on parameters input by the user through the HTML interface. The end-goal of ClusterSuite is to be an efficient, user-friendly application to assist in utilizing terms and phrases that were previously unrelated.