P-116

FLEXIBLE CHARACTERIZATION OF CARTOGRAPHIC GENERALIZATION RESOURCES FOR A SELF-ORGANIZING ONLINE CATALOG

WENDEL J., BUTTENFIELD B.P.
University of Colorado - Boulder, BOULDER, UNITED STATES

BACKGROUND AND OBJECTIVES

When searching online for published work or algorithms in the field of cartographic generalization it is often hard to identify and locate resources, which differ in format and media, ranging from journal articles, slide presentations, and algorithms to finished software. No centralized cataloging or indexing service is available. We are implementing an online, self-organizing catalog of generalization resources including entries in all formats listed above. This catalog will be available for users to contribute new entries. The catalog is based on Kohonen's Self-Organizing Map (SOM) algorithm. The presentation will cover the flexible characterization of multi-format resources for cartographic generalization, where the use of explicit keywords is not always possible. A pilot corpus of resources will be demonstrated which contains examples of full-text documents, published algorithms, datasets and software code.

APPROACH AND METHODS

SOMs are one type of artificial neural network. SOMs classify data to organize and clarify the relationships between data items, which make them beneficial for data searching and exploration. In a SOM, relationships between data items can be understood using a metaphor of distance as a measure of similarity, wherein similar data items are situated in close proximity and dissimilar items are situated farther apart. Using a SOM to organize a catalog or index to an archive holding items of diverse types permits the indexed items to be searched and retrieved efficiently regardless of how they are actually stored.

A major challenge in this research is to design a keyword characterization scheme which has three primary characteristics: a lack of redundancy; flexibility; and implicit keywords. The first characteristic, redundancy reduces efficiency of classification, thus the set of keywords should be independent of each other. This is a balancing act however, since using too many keywords will result in a categorization scheme which has too few examples for individual keywords, which results in a SOM network which is too sparse. Redundancy is therefore related to coherence. The second characteristic is flexibility, which is important to accommodate multiple layers of characterization that permit newly entered information to contribute to the organization of the catalog as a whole. The third characteristic relates to implicit keywords and this becomes especially important for categorizing resources in multiple formats. For example, the contents of a full-text journal article may be directly accessed to select explicit keywords, following a similar process to the selection of keywords listed at the beginning of this abstract. However, algorithms, datasets and software do not contain discriminatory keywords explicitly; as a consequence, implicit descriptor sets must be established. For example, datasets might contain raster or vector data (or both) so one implicit keyword might be “raster-only” or “raster-vector”. Algorithms might modify spatial relationships or attributes, or both, offering another category of implicit keywords. The trick is to create a set of implicit keywords which categorize various aspects of the corpus exhaustively, coherently and non-redundantly. Examples of explicit and implicit keyword sets relevant to various data formats are shown in Figure 1. Metadata, tutorials and help files provide common sources for the generation of implicit keyword sets (Wendel et. al. 2008).

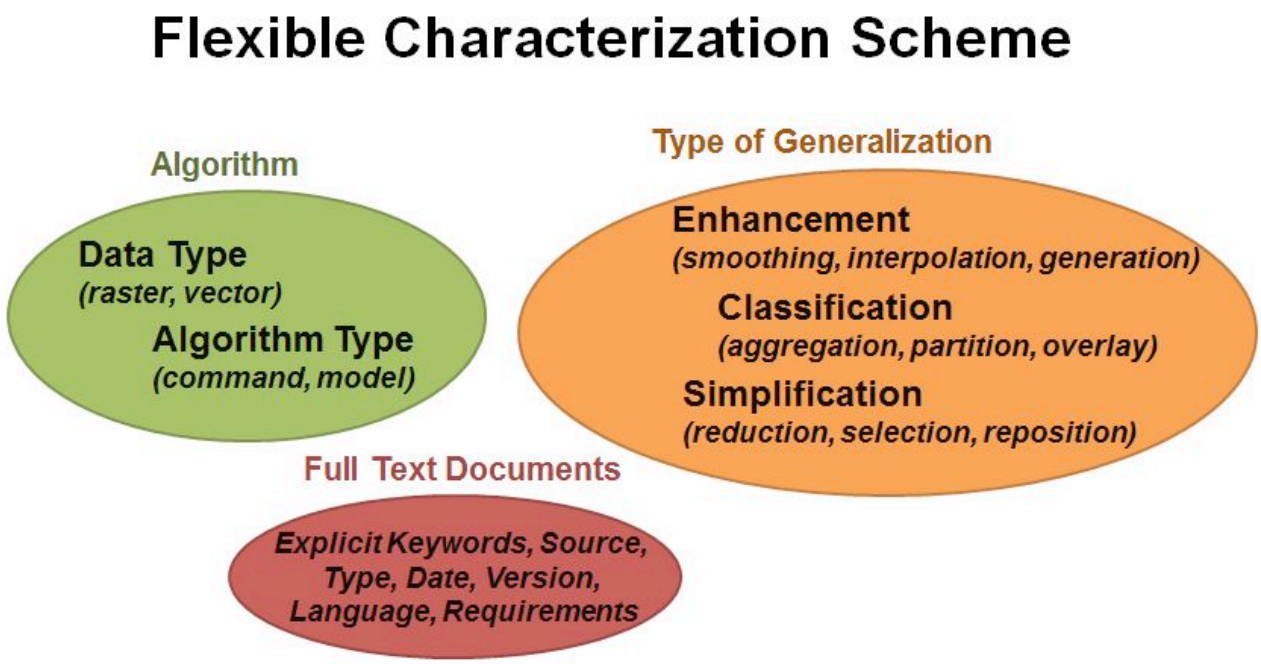

Figure 1. A set of implicit and explicit keywords which are coherent, flexible and exhaustive is needed to discriminate among multiple input formats. This figure shows a hierarchy of flexible keyword characterizations which can be applied to multiple input formats. This is not an exhaustive list, but it illustrates the range of types of descriptors needed to create a SOM. The keywords in bold fulfill minimum characterization requirements whereas words in italic are optional keywords to enhance the categorization.

FLEXIBLE CHARACTERIZATION AND EXISTING CLASSIFICATION SCHEMES FOR GENERALIZATION

In creating the SOM of generalization resources, we do not intend to build a new taxonomy of generalization procedures, but rather to make use of existing classification schemes which can be combined and extended to create a flexible multi-format classification strategy suitable for SOM operation. We also do not intend to build a data warehouse, rather to build a catalog pointing to online resources to improve accessibility to those resources for the generalization community.

RESEARCH METHODS

The presentation will discuss the sources and types of items to be categorized, the derivation of implicit keywords, SOM implementation and results. The test dataset used for this work will include a sample of generalization resources currently available on the Internet which are variously formatted as full text documents, software code, and datasets. Special emphasis will be given to the establishment of implicit keyword sets and their impact on the self-organization process, exploring for example what is the impact of incomplete characterizations on the final clustering.

ACKNOWLEDGMENTS

This research is supported by USGS grant # 04121HS029 "Generalization and Data Modeling for New Generation Topographic Mapping".