Adopting Mental Similarity Notions of Categorical Data Objects to Algorithmic Similarity Functions

David Sessler, Jürgen Bernard, Arjan Kuijper, and Jörn Kohlhammer

Motivation

- Similarity functions are essential for many analytical tasks
- Goal: create a similarity function based on visual-interactive user feedback to capture the Mental Similarity Notion in the heads of domain experts [1]
- Inspiring solutions for numerical data exist [2, 3]
- Problem: user feedback for categorical data attributes poses additional challenges (see Test Case)
- Our Solution: a Feedback Model with an additional improvement step for categorical attributes

Background

Illustrative example for similarity function definition for a countries data set [1]. Left: a user has arranged 15 objects which define the geographic topology of Europe. Right: the Result View represents the relations of countries of the whole data set with respect to the calculated similarity function (approximation of the geographic topology of the world).

The Feedback Model maps the 2D arrangement of data objects to pairwise object distances. Hence, the Feedback Model calculates attribute weights which are utilized to learn similarity functions. A visual representation completes the sense-making loop.

Preliminary User Study

Goal: Improvement of the Feedback Model on the basis of the user intuition
- Grasp the users intuition how to arrange objects in 2D
- Degree of freedom: number of categories (from 3 to 7)
- Number of participants: 11 (from academic fields)

Results representing the two most intuitive arrangements for 3 to 7 categories.

Insight: Users prefer object arrangements of regular geometry

Test Case

Basic Assumption:
- The weights of the categorical attributes should reflect the Mental Similarity Notion of the user
- The Feedback Model should generate perfect weights for the preferred arrangement geometries (the median weights of the empirical tests should be one)

Test Setup:
- Independent variable: feedback geometry (cardinality from 3 to 7)
- Independent variable: number of feedback objects (can be higher than the cardinality → stacking)
- Dependent variable: attribute weighting (0.0-1.0)
- 10000 tests for each configuration

Insights:
- Geometries with higher cardinality generate lower weights (x-axis)
  - Only the equilateral triangle (left) reaches the weight of 1.0 (pairwise distances identical)
  - Improvement strategy needed for other geometries

- For each geometry the weighting increases with the number of feedback objects (nested x-axis)
  - needs to be considered in the improvement step

- User-preferred geometries perform better than the second choice geometries
- Feedback Model complies with user intuition

Improvement Strategy:
Use the median weights of the user-preferred geometries to improve the attribute weight generation
- 1. Set the obtained median weights as new maximum
- 2. Rescale the weighting result in the range of [0.0-1.0]

Expected outcome for a Test Repetition:
Median weightings of the preferred geometries should reach the weight 1.0

Results of the repetition of the empirical test under same conditions with an incorporation of the improvement step.

Result of the Test Repetition:
- The median weightings are 1.0 for each user-preferred geometry
- The median weightings are near 1.0 for the second choice geometries
- The weightings of the categorical attributes reflect the Mental Similarity Notion of the user

Future Work

- Test the dependency of generated attribute weightings on non-uniform distributed attributes
- Test improved Feedback Model for Mental Similarity Notions based on more than one attribute
- Apply improved Feedback Model in a mixed data set. Test interplay with numerical attributes
- Test the improved Feedback Model in a real world example

References