UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Latent Semantic Analysis Approaches to Categorization

Permalink

https://escholarship.org/uc/item/1hb5j83q

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 19(0)

Author

Laham, Darrell

Publication Date

1997

Peer reviewed

Latent Semantic Analysis Approaches to Categorization

Darrell Laham
Department of Psychology & Institute of Cognitive Science
University of Colorado, Boulder
Boulder, CO 80309-0345
dlaham@psych.colorado.edu

Many computational models of semantic memory rely on vector representations of concepts based on explicit encoding of arbitrary feature sets. Latent Semantic Analysis (LSA) creates high dimensional ($n = 300+$) vectors for concepts in semantic memory through statistical analysis of a large representative corpus of text rather than subjective feature sets linked to object names (for details see Landauer & Dumais, 1997; Landauer, Foltz, & Laham, in press). Concepts can be compared in the semantic space and their similarity indexed by the cosine of the angle between vectors.

Computational models of concept relations using LSA representations demonstrate that categories can be emergent and self-organizing based exclusively on the way language is used in the corpus without explicit hand-coding of category membership or semantic features. LSA categorization is context dependent and occurs through a dynamic process of induction. Semantic "meaning" is not encapsulated within an object representation, but emerges as the set of relationships between selected objects in a context-based sub-space.

Neuropsychological studies (e.g. Warrington & Shallice, 1984) point to a class of patients who exhibit disnomias for specific categories of objects (natural kinds) while retaining the ability to name other objects (man-made artifacts). The objects from natural kind categories tend to be significantly more clustered in LSA space than are those from artifact categories. If brain structure corresponds to LSA structure, the identification of concepts belonging to strongly clustered categories should suffer more than weakly clustered concepts when their representations are partially damaged.

Three types of modeling experiments were conducted: matching base concept names to superordinate categories in forced-choice testing, correlating LSA similarity measures to human judgments of typicality, and multivariate analyses of similarity matrices to capture category boundaries.

For the forced-choice matching of concept names to superordinate categories, a selection of 140 objects (rated as most typical in their category) from 14 categories was used. Each object name was compared to each of the 14 category names (apple—flower, apple—mammal, etc.). The LSA match was considered correct when the highest cosine comparison in the set was between an object and its relevant superordinate (apple—fruit). The results show that in all 14 categories, LSA predicts membership well above chance (chance = 7%), however, there are differences in the degree of clustering: the percent correct for animate natural kinds (flowers, mammals, fruit, trees, vegetables, and birds) = 92%; for inanimate natural kinds with observed deficits in neuropsychological patients (gemstones, musical instr-

uments) = 100%; and for man-made artifacts (furniture, vehicles, weapons, tools, toys, and clothing) = 53%.

Correlations between LSA similarity judgments and human typicality judgments were consistently better for the natural kinds than for the artifacts. For natural categories, LSA similarities (cosine between concept and either superordinate name, most typical member, or centroid of all members) showed high correlations with human judgments (e.g. fruit: $r = .82$), while artifact similarities showed low to near-zero correlations with human judgments.

As illustrated in Figure 1, multivariate analyses of LSA-based similarity matrices show more cohesive structure for natural kinds than for artifacts. Factors 4-6 in this analysis load high on concepts in the bird category—additional factors (7-15) load on specific artifact concepts (not shown).
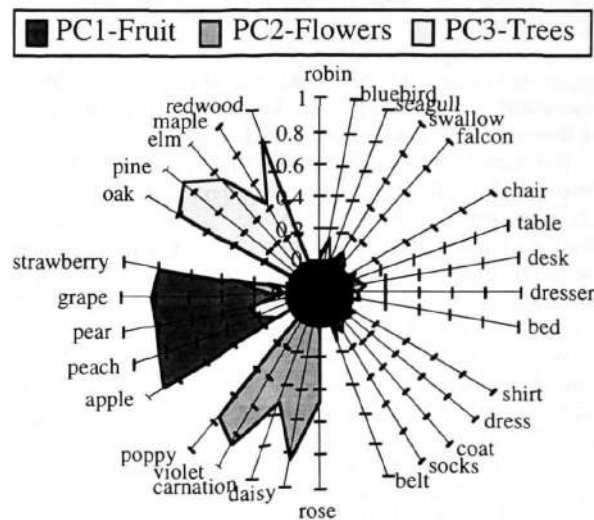

Figure 1. Principal Components (PC) Factor Loadings

References

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (in press). Introduction to Latent Semantic Analysis. *Discourse Processes*.

Warrington, E. K., & Shallice, T. (1984). Category-specific semantic impairments. *Brain*, **107**, 829-853.