Dubin, D. (1995). The Search for Structure and the Search for Meaning. 6th ASIS SIG/CR Classification Research Workshop, 13-20. doi:10.7152/acro.v6i1.12658

The Search for Structure and the Search for Meaning

David Dubin
Department of Information Science,
University of Pittsburgh
dsdst3@lis.pitt.edu

Statistical approaches to classification emphasize apprehension of structure by an analyst in a group of records, but issues of meaning and semantics are important, despite the focus on structure and algorithms. If meaning and semantics guide formal approaches to classification, can an understanding of structure in a collection of records inform the development of a semantic classification scheme? Data visualization tools can help human analysts recognize structure and pattern in text and numeric data.

1. Introduction[1]

A classification system is a set of constructs used by one or more persons to bring sense and structure to some part of reality. One often draws the distinction between "formal" systems with explicit structure (e.g. a library classification) and "informal" systems such as an ethnobiological classification or a classification of kinship terms (Eastman and Carter, 1994). One can make a further distinction between statistically-based and semantically-based formal systems. In a semantic system, the creation of category structure and the assignment of entities to categories is achieved through human intellectual effort. Statistical approaches involve applying an algorithm to representations of entities on one or more quantified attributes. The analyst's goal may be a category structure that fits the observations (e.g., cluster analysis) or a function that helps discriminate records in an existing category structure (e.g., discriminant function analysis) (Manly, 1986).

Classification by minds and classification by machines would seem to be at opposite poles, but issues of structure and of meaning are at the heart of both approaches. Automatic clustering methods support the goal of revealing structure, but users of those methods are cautioned to be guided by theory. Can the apprehension of structure in a collection of records inform the development of a semantic classification scheme? Data visualization tools, such as those under development at the University of Pittsburgh, may assist human classifiers by highlighting the same statistical properties to which clustering methods are sensitive.

2. Cluster Analysis

In cluster analysis, a collection of **n** representations is divided into **m** groups based on estimated or computed associations between pairs of representations. For example, one can crudely estimate topical similarity between documents on the basis of terms they share (Salton, 1989). A computer program can extract the terms and quantify them in document representations with a term weighting scheme. One can compute similarity between texts using a measure such as Euclidean distance or the cosine measure (Dubin, 1994).

[1] I'm grateful to Barbara Kwasnik and to Robert Korfhage for their comments on earlier drafts of this paper.

ISSN: 2324-9773

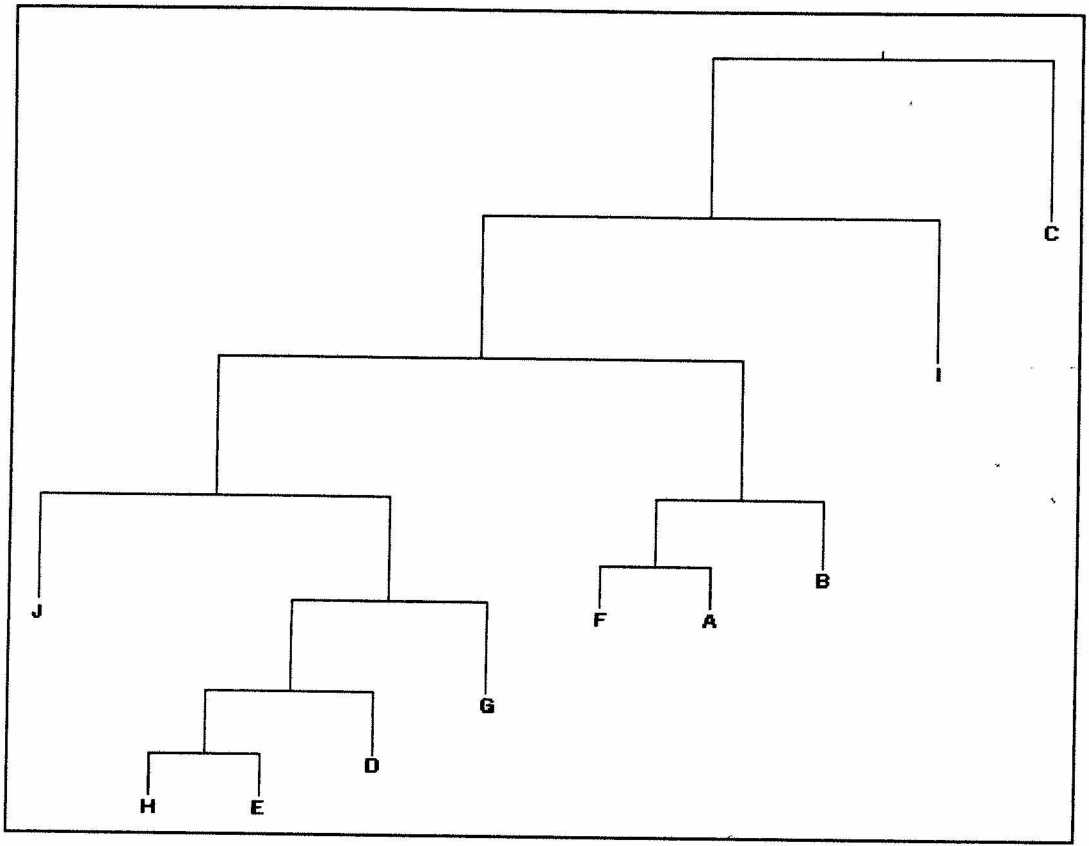

Figure 1: Dendrogram

A clustering algorithm groups representations into categories so that entities in a cluster are similar to each other but dissimilar to entities in other clusters. This can be accomplished either by successively grouping entities into larger and larger clusters, or by partitioning the data into groups (Aldenderfer and Blashfield, 1984). The former (hierarchical) family of methods includes the single, complete, and group average link methods. Nonhierarchical algorithms include single pass and reallocation methods (Rasmussen, 1992). Hierarchical clustering methods usually yield output in the form of a dendrogram. A dendrogram representing a hypothetical clustering is shown in figure 1. Dendrograms indicate the level of computed similarity at which two or more records are counted in the same cluster; this highlights no only the similarity relationships between records within a cluster, but also relationships among clusters themselves.

Both structural and semantic issues are relevant to decisions made during a cluster analysis. For example, before applying similarity measures to a set of records, the analyst must select the variables (attributes) which will participate in the calculation of similarity. These decisions are at a critical intersection of structure and semantics: the choice of attributes define what it is in the entities that the analyst considers relevant to their classification. Variables are

ISSN: 2324-9773

Dubin, D. (1995). The Search for Structure and the Search for Meaning. 6th ASIS SIG/CR Classification Research Workshop, 13-20. doi:10.7152/acro.v6i1.12658

representations or encodings of attributes understood by humans. If an attribute is quantitative (e.g., mass or age) then the encoding process is straightforward. But an analyst may encode a qualitative attribute with a quantitative approximation (e.g., IQ test score as a measure of intelligence). Alternatively, entities may be represented as binary feature vectors, where the presence or absence of a feature is encoded with a one or a zero, respectively.

Aldenderfer and Blashfield (1984) discuss the variable selection issue from a semantics perspective, arguing that the selection of attributes should be guided by an explicit theory. The theory, they state, should inform the selection of a set of variables that "best represents the concept of similarity under which the study operates" (page 20). The point of this advice is to discourage "naive empiricism" (i.e. the belief that structure will emerge from a large number of indiscriminately chosen variables). Clustering algorithms will always group representations, regardless of whether the data have a clustered, uniform, or random structure. The dendrogram in figure 1, for example, is the result of applying a clustering algorithm to ten points scattered randomly in three-dimensional space. Selecting variables in the context of a theory reduces the danger that a grouping may be imposed on the data by the method instead of by the meaning.

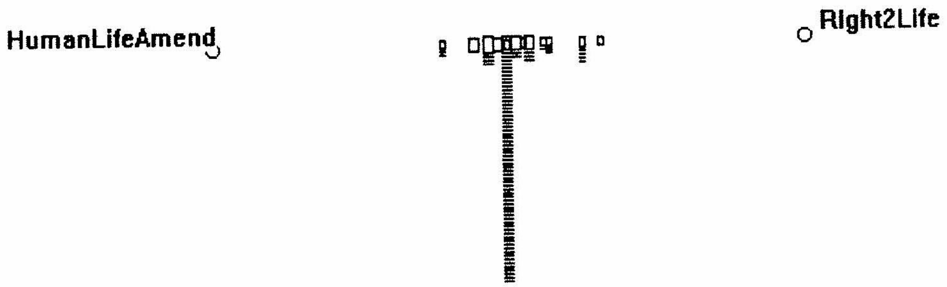

Figure 2: Effect of Positive Correlation

The same issue of variable selection is examined from a structural perspective in an article by Milligan (1989). Milligan agrees that it's a bad idea to select variables carelessly, stating that only those that "contribute or define the clusters that may exist" should be included. From the structural perspective, it is assumed that a meaningful clustering exists in the data, that "relevant" variables will highlight that clustering, and that irrelevant variables will "mask" or interfere with the recovery of that structure. Milligan goes on to cite evidence of the effect of masking variables, and to recommend automatic techniques that may reduce those effects.

3. Semantic Approaches: a Role for Visualization?

Although quantitative approaches to classification are very popular, they're not for everyone. Two objections to the application of cluster analysis in a study or application are that the variables of interest may be qualitative rather than quantitative, and that human minds may be better suited to judging similarity than (less subtle) association measures. When classifying

ISSN: 2324-9773

documents by topic, for example, one can get a pretty fair indication of a text's subject matter by extracting terms on the basis of frequency and assigning a numeric weight based on term occurrence (Salton, 1989). But those approximations are extremely crude in comparison to a trained indexer's application of intellectual effort. Similarity measures are sensitive to components of variability that can help group related documents (Jones and Furnas, 1987), but humans are capable of far more subtle judgments.

The problems with clustering, then, would seem to be that quantitative approximations can't always *replace* qualitative judgment, and that mechanical assignment of entities to groups isn't always the most helpful way to reduce data complexity. But it's possible that an analyst classifying on the basis of meaning might explore a data set and generate hypotheses with the aid of quantitative approximations. Such an analyst would need tools that highlight structure, while letting the exploration be guided by human judgments of meaning.

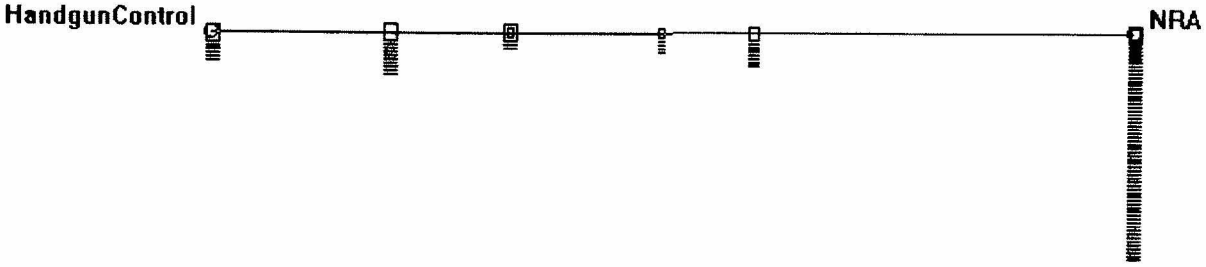

Figure 3: Effect of Negative Correlation

Figures 2, 3, 4 and 5 show how that kind of tool might be used. The figures are screens taken from VIBE, a program developed at the University of Pittsburgh and Molde College for visually exploring multivariate data (Olsen et al., 1993). The figures show a data set derived from Project Vote Smart's performance evaluation information for 100 Republican representatives. Project Vote Smart is a project of the Center for National Independence in Politics. The voting evaluations track the percentage of an incumbent's voting record consistent with the preferred positions of various special interest groups.

VIBE is one of several related visualization tools under development at Pitt's Department of Information Science (Olsen and Korfhage, 1994) (Kim and Korfhage, 1994) (Nuchprayoon and Korfhage, 1994). In VIBE, records are represented by rectangles, and are plotted on the screen according to the ratios of the scores they receive on quantified attributes (represented by circles). VIBE plots similar records near each other, and identical records in "stacks" represented by vertical bars (as seen in the figures). Individual and stacked record icons will be closest icons representing attributes with the highest numeric scores. The attribute icons (called POIs, or Points of Interest) are freely positionable on the display, and can be activated and deactivated by the analyst at will. The analyst can view the data set from the perspective of two or more attributes, and see how those attributes group or separate the

ISSN: 2324-9773

records. He or she can then select records with the mouse, and view the raw data on all scored attributes. POI influence on an individual record icon can also be revealed with a "star" display, as shown on the left side of figure 4.

VIBE can be used to visualize associations in both text and numeric data. The plotting function, using ratios of attribute scores, is sensitive to the same components of variability as angular similarity measures (e.g., the cosine measure). However, VIBE does not assign records to clusters or categories.

Suppose that an analyst is interested in classifying House Republicans according their positions on key issues. One can view the assessments of special interest groups as crude quantitative approximations to the representative's positions on those issues. The assessments are at least one step removed from an actual voting record, which is at least one step removed from the representative's positions. Visualizing the data with a tool like VIBE puts the analyst another step away, since it shows information on relative rather than absolute scores. What useful information can VIBE's display provide to the analyst developing a classification scheme?

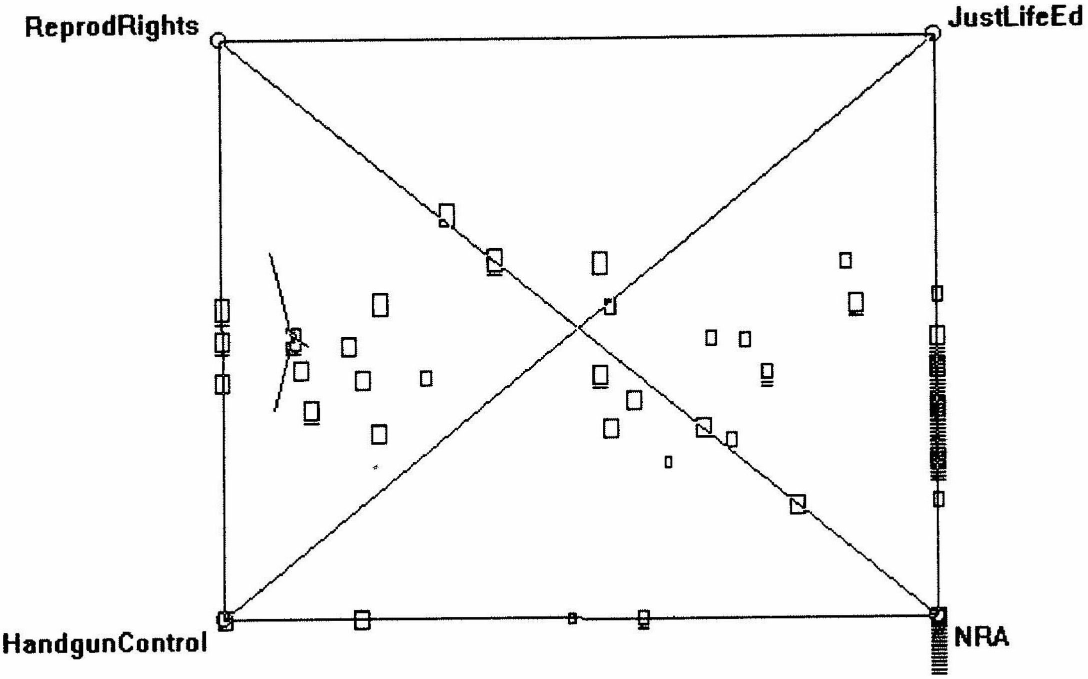

Figure 4: Strong Discriminators

A classification scheme should highlight contrasts and differences: it's no good to develop a scheme if all entities are assigned to the same category. VIBE could help an analyst identify

ISSN: 2324-9773

Dubin, D. (1995). The Search for Structure and the Search for Meaning. 6th ASIS SIG/CR Classification Research Workshop, 13-20. doi:10.7152/acro.v6i1.12658

the issues that divide House Republicans by displaying the effect that different variables have in literally pulling apart the icons on the display. Statistical properties like variability in assigned scores and dependencies between variables can inform the selection of variables, but interpreting those properties and their effects requires understanding and judgment by humans. Figures 2 and 3 illustrate the effect of dependencies between variables on their power to discriminate. Figure 2 plots the records according to assessments of the National Committee for a Human Life Amendment and the National Right to Life Committee. This plot shows no discrimination at all, since scores for those two groups will be positively correlated. In figure 3 the records are plotted according to assessments of the National Rifle Association and Handgun Control, Inc. These negatively correlated variables discriminate clusters on both sides of the issue, as well as four clusters of representatives with moderate voting records. Computer programs can help identify sets of variables that are highly intercorrelated, and the understanding of human analysts can inform decisions as to which of the sets to use in the study or application.

Figures 4 and 5 illustrate the role that score variability can have on the power of individual variables to discriminate records. In figure 4, two additional variables have been added to those in figure 3: assessments of the Justlife Education Fund and of the National Abortion Reproductive Rights Action League. In addition to a negative correlation, scores on these two attributes are more variable (as measured by standard deviation) than the variables chosen for figure 5. Explaining the difference in variability requires investigation and interpretation by humans: it may be that one set of issues is more divisive, or simply that groups like the Christian Coalition and the American Civil Liberties Union assess records on a larger number of different issues than the gun control and abortion groups.

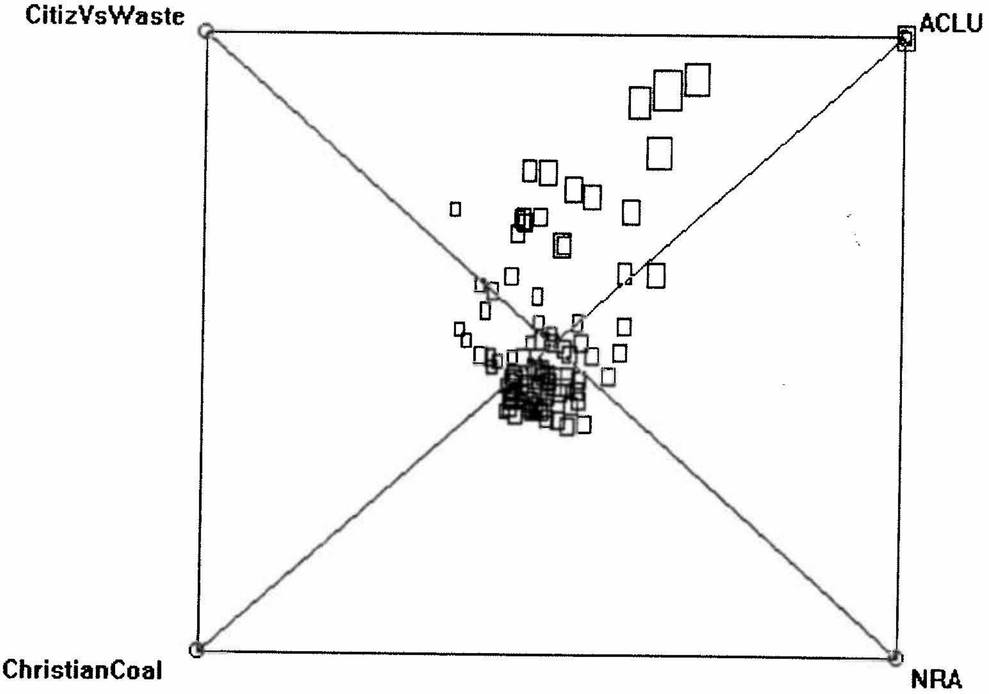

Figure 5: Weak Discriminators

ISSN: 2324-9773

Conclusions

A strategy for using formal properties to select variables for record classification and retrieval tasks in currently under development and evaluation (Dubin, 1995). With this strategy, automated methods call analysts' attention to formal properties such as the variability and intercorrelations. But analysts play an active role in identifying the meaningful variables, and proposing combinations that make sense in the context of the study or application. In this way, minds and machines each contribute information they are best suited to recognize.

ISSN: 2324-9773

References

Aldenderfer, M. S. and R. K. Blashfield. (1984). *Cluster Analysis*. Sage Publications.

Dubin, D. (1994). Applying Similarity Measures to Texts. *TEXT Technology*, 4(4), 283-291.

Dubin, D. (1995). Document Analysis for Visualization. *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. 199-204.

Eastman, C. M. and R. M. Carter. (1994). Anthropological Perspectives on Classification Systems. *Proceedings of the 5th ASIS SIG/CR Classification Research Workshop*.,69-77.

Jones, W.P. and G. W. Furnas. (1987). Pictures of Relevance: A Geometric Analysis of Similarity Measures. *Journal of the American Society for Information Science*, 38(6), 420-443.

Kim, H. and R. R. Korfhage. (1994). BIRD, A Browsing Interface for the Retrieval of Documents. *Proceedings of the 1994 IEEE Symposium on Visual Languages*. 176-177.

Manly, B. F. J. (1986). Multivariate Statistical Methods: A Primer. Chapman & Hall.

Milligan, G. (1989). A Validation Study of a Variable Weighting Algorithm for Cluster Analysis. *Journal of Classification*, 6, 53-71.

Nuchprayoon, A. and R. R. Korfhage. (1994). GUIDO, a Tool for Retrieving Documents. *Proceedings of the 1994 IEEE Symposium on Visual Languages*. 64-71.

Olsen, K.A. and R. R. Korfhage. (1994). Desktop Visualization. *Proceedings of the 1994 IEEE Symposium on Visual Languages*. 239-244.

Olsen, K. A., R. R. Korfhage, K. M. Sochats, M. B. Spring and J. G. Williams. (1993). Visualization of a Document Collection: The Vibe System. *Information Processing and Management*, 29(1), 69-81.

Rasmussen, E. (1992). Clustering Algorithms. In Frakes, W. B. and R. Baeza-Yates (Eds.), *Information Retrieval: Data Structures and Algorithms*. 419-442, Prentice Hall.

Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley.

ISSN: 2324-9773