Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

Computer-aided Interactive Classification: Applications of VIBE

Robert R. Korfhage
Dept. Information Science
University of Pittsburgh
korfhage@lis.pitt.edu

David S. Dubin
Graduate School of LIS
University of Illinois
dubin@alexia.lis.uiuc.edu

Abstract

Tools like the VIBE visualization system permit human analysts to use both an understanding of a data set's content and a recognition of structure that the visualization reveals. But what happens when a database's semantics are hidden from the analyst? What guidelines or heuristics can he or she use to reveal the "correct" underlying structure? Results of two experiments conducted at the University of Pittsburgh support the claim that VIBE analysts can uncover a meaningful clustering even without semantic clues. In one experiment artificial data sets were created in which some of the variables discriminate one or more clusters and the other half contribute only random noise. Variable selection guidelines based on computed discrimination value were used in an attempt to distinguish between the signal and noise variables. In a second experiment, a human analyst's encoding of 714 short phrases to 23 overlapping and inter-related categories was stripped of meaningful titles and relabeled with integers. A VIBE analyst was able to highlight relationships among the 23 categories solely on the basis of co-assignment of the phrases.

1 Introduction

One argument for the use of visualization interfaces in classification is that they can make explicit the relationships among documents and the structure of a document set with a clarity that cannot be achieved using conventional user interfaces. This claim raises the interesting question of the source of that structure: how much is inherent in the document set, how much is imposed by the user's semantics, and how much is an artifact of the interface itself. When human beings classify records the organization originates in the mind of the classifier, and records are assigned to categories based on an analyst's understanding of their content. When records are classified by an algorithm (such as a clustering procedure), the classification is supposed to emerge from the structure of the records. Although semantic

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

and automatic approaches to classifying differ, users of both methods are concerned with the structure and meaning of records. Visualization interfaces offer analysts opportunities to apprehend structure in a set of records that can inform or guide their assignment to categories. The procedures that plot record icons in a display are sensitive to the same kinds of variability as clustering algorithms, but such interfaces need not assign records to categories.

The rationale for computer-aided classification is that it offers advantages of both the automatic and semantic approaches: decisions about how a record should be classified aren't delegated to a machine, and the perception of pattern in a matrix of data is made easier with a tool. But a number of new problems arise, including questions of the validity or reality of the patterns perceived in the visual display. In purely semantic approaches to classification human beings bear responsibility for the appropriateness of a record's assignment. The technology of cluster analysis includes tools for ruling out the possibility that an organization has been imposed upon random data [Dubes & Jain, 1979]. But how does one know with a visualization of multi-dimensional data whether one can believe what one sees? What does it mean for patterns to be "true" or "false?" This paper reports two experiments intended to address these issues.

2 The VIBE System

Experiments reported in the following sections employed VIBE, a system for graphically depicting similarity relationships among documents [Olsen et al., 1993]. VIBE users can specify an arbitrary number of points of interest (POIs) or reference points, and freely position icons representing these POIs on the computer screen. The POIs can be individual terms, complex queries, or known documents. All documents in the document set that relate to these POIs will then be displayed as icons positioned according to the strength of the relationships between the POIs and the documents. While VIBE is fundamentally a vector-based system, a Boolean version of it exists. Under VIBE's vector display, rectangular icons represent documents, and circular icons represent document attributes selected as points of interest. VIBE's plotting function is equivalent projection of document vectors on to the surface of a k-dimensional hypersphere defined in the L1 (or city block) metric. This means that documents having similar ratios of POI weights (e.g., documents similar according to the cosine measure) are plotted near each other on the VIBE display. L1 lengths of document vectors do not participate in the positions of document icons, but do control their size. Under the Boolean display, records are grouped according to the combinations of POIs influencing them. Integers printed in rectangles indicate how many records satisfy a particular Boolean combination of variables.

3 Imposed vs. Discovered Structure

One hopes that the analyst assisted by VIBE is perceiving real structure in the data, and that the display is not merely a Rorschach blot. There are at least three ways in which VIBE's display could be misleading:

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

1. VIBE's positioning function for the document icons is based on both the POI weights (i.e., attribute scores) and the positions of the movable POI icons. Projection of a high-dimensional vector space onto a two-dimensional display inevitably results in distortion of associations in the original data. In VIBE it's possible for two record icons to be plotted near each other, even though the records have no attributes in common.

2. VIBE can offer the analyst little help in selecting record attributes for visualization. POIs with randomly assigned scores can often be distinguished from discriminating POIs by the visual scattering effect that they have on the display. However, a trial and error approach for POI selection is unworkable in applications such as text analysis where the number of potential attributes is very large and the number of attribute combinations explodes combinatorially. Furthermore the presence of randomly distributed attributes adds enough visual noise to VIBE that structure discriminated by other POIs can become invisible. Some kind of filter is desirable to help distinguish "signal" from "noise" attributes before the analysis begins.

3. Ultimately VIBE's usefulness depends on whether the structure it makes explicit is meaningful and interesting in the context of the application. At best, the success rate will vary from application to application, and structure in a data set that VIBE cannot show might be revealed by some other visualization tool. That VIBE's icon positioning is based on a simple vector sum makes it easier to predict associations it can and cannot show[1]. Nevertheless, a very compelling question is whether VIBE can help detect interesting structure when few or no semantic clues are available to the analyst.

The problem of distortion due to projection of the high-dimensional space is ameliorated by VIBE's interactivity (e.g., the movable POIs) [Olsen & Korfhage, 1994]. It's true that any given POI configuration gives an incomplete picture of the multidimensional space, but an experienced VIBE user can manipulate VIBE's display and test hypotheses about perceived structure.

For example, the apparent influence of a POI on a record icon can be tested by moving the POI or by highlighting all influenced records with a color. Icons for group of similar records will remain near each other, even when the POIs influencing them are repositioned. Armed with an understanding of VIBE's plotting function, the analyst can make sure his or her inferences about the records are correct.

Experiments reported in the following sections are aimed at the problems of variable selection and analysis in the absence of semantic clues.

4 Attribute Selection

An investigation of the document attribute selection problem for tools like VIBE has been reported in an earlier paper [Dubin, 1995]. In that study measures of term co-occurrence

[1] With systems based on methods like simulated annealing [Chalmers & Chitson, 1992] and neural networks [Lin, 1992] it's more challenging to explain how exactly icons come to be plotted where they are.

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

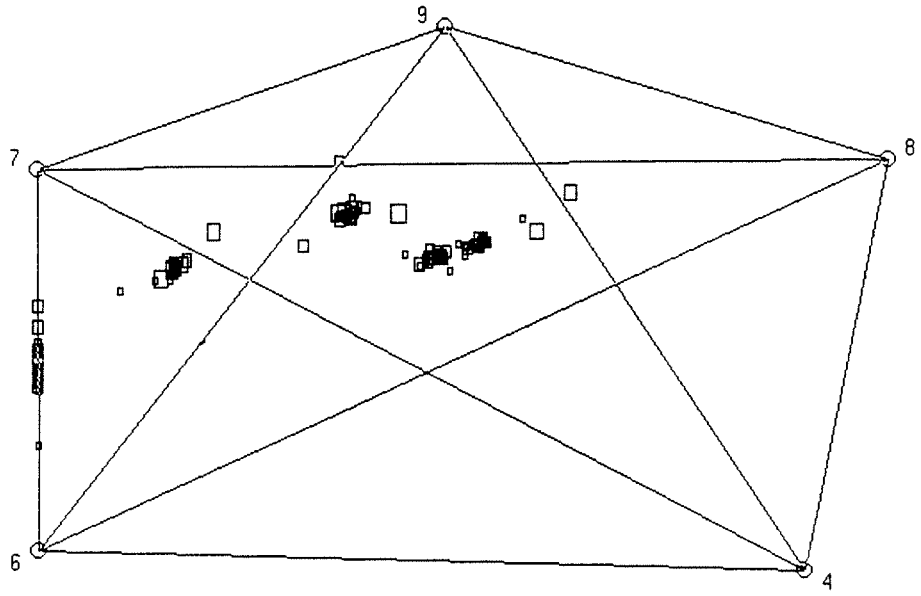

Figure 1: Five of the ten discriminating dimensions

and computed discrimination value of extracted stems guided choice of reference points. Stems with the highest term discrimination values [Salton et al., 1975] and stem sets with the least overlap defined vector spaces with the greatest tendency to cluster (as measured by two diagnostics adapted from the field of cluster analysis).

Early experiments with the attribute selection guidelines indicated that one can distinguish those attributes likely to reveal clustering in document spaces from those unlikely to reveal such structure. However, the experiments offered no evidence that the guidelines could help uncover any particular clustering (e.g. one known in advance to exist). A more recent and thorough evaluation of the attribute selection guidelines has included a test of their ability to recover known structure from a set of artificial data [Dubin, 1996]. Some results of this ongoing investigation are reported below.

These experiments employ data generated with an algorithm written by Glenn Milligan as part of his research in automatic variable weighting for cluster analysis [Milligan, 1985]. The algorithm defines centroids and boundaries for clusters in a multidimensional space, and then populates the clusters with vectors drawn from a multivariate normal distribution. Various types of noise (including outliers and dimensions consisting of random data) are added to the data before it is written to an output device. In the current study, Milligan's algorithm was modified so that clusters were defined on one or more (but not necessarily all) of the discriminating dimensions [Dubin, 1996]. Figures 1 and 2 show two examples of the multidimensional data projected into the plane of VIBE's display. In figure 1, dimensions on which a cluster is not defined receive null values for that cluster. Figure 2 shows data to which random noise has been added to all discriminating dimensions, whether or not a particular cluster is defined on that dimension.

Two data sets were created using the algorithm: each consisting of some discriminating dimensions and some random noise dimensions. One data set (see table 1) consisted of

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

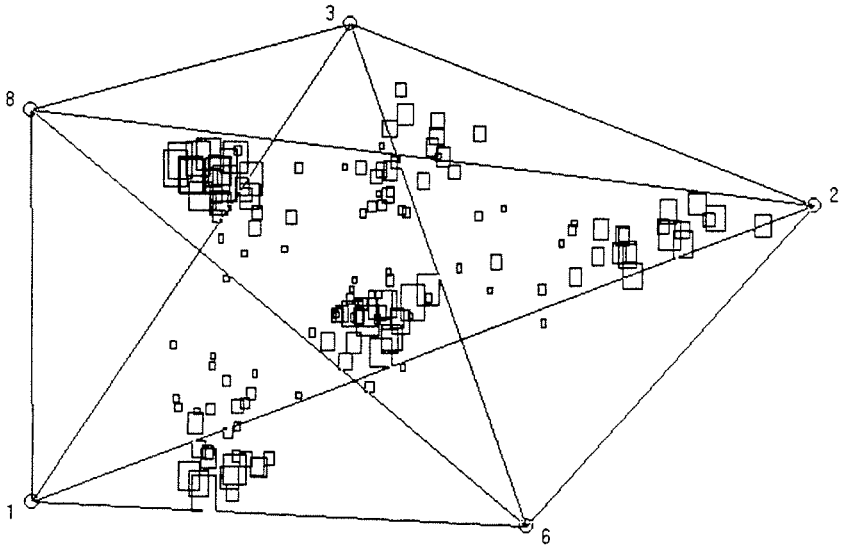

Figure 2: Five of eight discriminating dimensions (noise added)

	1	2	3	4	5	6	7	8
Cluster 1		X		X	X			X
Cluster 2		X		X	X		X	
Cluster 3	X			X		X	X	
Cluster 4		X	X	X			X	
Cluster 5	X		X			X	X	

Table 1: Dimensions discriminating clusters in twelve-dimensional data

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

	1	2	3	4	5	6	7	8	9	10
Cluster 1			X		X	X	X		X	
Cluster 2	X		X			X		X	X	
Cluster 3		X				X		X	X	X
Cluster 4	X					X	X	X	X	
Cluster 5		X			X	X	X			X

Table 2: Dimensions discriminating clusters in twenty-dimensional data

4 noise dimensions and 8 discriminating dimensions that defined 5 clusters of 30 records each. Each cluster was defined on four of the 8 discriminating dimensions (as shown in the table), and random noise was added to all 8 of those dimensions. The second set consisted of 10 noise dimensions and 10 discriminating dimensions that defined 5 clusters of 24 records each. Each cluster was defined on 5 of the 10 discriminating dimensions (table 2, but no additional noise was added to that data beyond the 10 noise dimensions (figure 1).

In earlier document analysis experiments, extracted stems were ranked by term discrimination value (TDV) as a guideline for selecting POIs likely to discriminate clusters. As described by Salton, a term's discrimination value is a measure of its contribution to average pairwise document similarity across the collection [Salton et al., 1975]. Average similarity is computed (or approximated by average similarity of each document to a collection centroid) with and without the term's participation, and the difference between those values is the TDV. Terms or stems with a positive discrimination value make documents seem less similar to each other on average, while terms with negative discrimination values increase average similarity.

The rationale for TDV's use as a selection guideline in these studies is that a term or attribute with a strong representation in a subset of the collection will distinguish that subset from the rest of the documents. The contrast along such a dimension between the subset characterized by the attribute and those not characterized should lower the average similarity among all documents when the discriminating attribute participates in the similarity computations. One therefore expects that attributes that strongly discriminate an identifiable cluster should have higher TDV values than attributes characterizing most or all documents in the collection. However, it's possible that a stem or other attribute can have a high TDV without discriminating a cluster: choosing an attribute with random but highly variable scores can just as easily decrease average similarity among documents.

For each data set, all dimensions (discriminating and noise) were ranked by discrimination value. Of interest was whether all discriminating dimensions would be ranked higher than the noise dimensions, whether all discriminating dimensions would receive a positive TDV, and whether all noise dimensions would receive a negative TDV.

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

4.1 Results

The 12 dimensions of the first data set were ranked by discrimination value. Dimensions 1-8 are the discriminating dimensions and 9-12 are the noise dimensions.

1. dimension 3: 0.00689
2. dimension 8: 0.00615
3. dimension 2: 0.00496
4. dimension 6: 0.00412
5. dimension 1: 0.00166
6. dimension 5: 0.00121
7. dimension 7: 0.00034
8. dimension 4: -0.00057
9. dimension 9: -0.00344
10. dimension 11: -0.00448
11. dimension 10: -0.00569
12. dimension 12: -0.00661

All 8 of the discriminating dimensions have a higher TDV than the noise dimensions. However, the lowest ranked of the discriminating dimensions (number 4) has a negative TDV: average document similarity was lower when the term was absent from the collection than when it was present. Note that the two lowest ranked signal dimensions (7 and 4) were those on which the largest number of clusters (and hence the majority of the records) are defined.

The 20 dimensions of the second data set were ranked by discrimination value. Dimensions 1-10 are the discriminating dimensions, and 11-20 are the noise dimensions.

1. dimension 1: 0.03639
2. dimension 2: 0.01612
3. dimension 10: 0.01185
4. dimension 3: 0.01156
5. dimension 5: 0.00992
6. dimension 8: 0.00559
7. dimension 7: 0.00537

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

8. dimension 4: 0.00000
9. dimension 18: -0.00154
10. dimension 16: -0.00310
11. dimension 19: -0.00359
12. dimension 17: -0.00400
13. dimension 12: -0.00474
14. dimension 14: -0.00508
15. dimension 20: -0.00513
16. dimension 15: -0.00513
17. dimension 11: -0.00556
18. dimension 13: -0.00967
19. dimension 9: -0.01211
20. dimension 6: -0.02629

Although dimension number 4 is one of the 10 original "signal" dimensions, the data generating algorithm did not choose it to discriminate any of the five clusters (table 2). All the scores along that dimension were null, and so the dimension has a TDV of zero. Seven of the remaining nine discriminating dimensions have positive discrimination values. However, dimensions 9 and 6 not only have negative TDV scores, but the values are lower than any of the noise dimensions. Note that four of the five clusters are defined on dimension 9 and all five clusters are defined on dimension 6.

4.2 Interpretation

Further experimentation is required with a variety of artificial data sets, but it's clear that computed TDV can't always distinguish dimensions that discriminate clusters from dimensions with randomly distributed scores. Our interpretation of these results is that Dimensions 7 and 4 in the twelve-dimensional set and Dimensions 9 and 6 in the twenty-dimensional set received low discrimination values because the majority of the records were defined along those dimensions. Although records were grouped into clusters along those dimensions, the resulting contrasts among the records weren't as significant as those produced by a highly variable, random assignment of scores. We're encouraged, however, to have had more success with the noisier data than with the less noisy data: making all the dimensions more variable may have helped highlight the difference in clustering. Current research is directed toward adapting the TDV measure to overcome its sensitivity to highly variable dimensions.

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

5 Analysis of Uninterpreted Data

This section reports an experiment in which data were coded by one analyst so that meaningful names were replaced by integers. The data were turned over to a second analyst who organized the records into groups with the aid of VIBE. Those groups were later reviewed by the first analyst in light of his understanding of the collections semantics.

5.1 The Documents

The documents used in this experiment were 714 short phrases, each describing a data element in a group of military databases under study at The MITRE Corporation. Examples of these phrases are:

- BOOSTER-IMPACT-POINT-LOCATION
- OBSERVED-OBJECT-BEARING-ANGLE

The phrases describe data concepts that are important to manage in a military operation where an enemy SCUD attack is anticipated.

Each of these phrases, which MITRE terms a "text object," was reviewed by an analyst at MITRE and assigned to one of 23 POIs categorizing the data. Of interest was whether POI groups could be organized into broader concepts that could be used to model the data. Examples of these POIs are:

- SENSOR MANAGEMENT
- WEATHER
- AIRSPACE CONTROL

Each text object and each POI were assigned integer identifiers at random; these identifiers constituted the data that were presented to the experimenters. Thus, the representation for a document might be (12: 3, 4, 15, 20, 21), where '12' is the text object number, and the remaining integers represent the relevant POIs. Because of preprocessing of the text objects, seven of the objects were related to none of the POIs, yielding an effective set of 707 documents for the experiment.

Since each POI is listed at most once for any text object, we could have used either the vector or the Boolean version of VIBE with equal results. The Boolean version was chosen because its visualization is somewhat better adapted to the task at hand. In the vector version, multiple text objects at a given point are represented by a 'tail' on the text object icon, with length proportional to the number of text objects; in the Boolean version an explicit text object count is given, rather than the tail.

5.2 POI Relationships

Co-assignment of text objects to POIs is at the heart of the structure revealed in this data. A pair of POIs might stand in any of several relationships to each other: they might

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

be completely independent (disjoint text object sets), one might completely subsume the other, or the sets of text objects assigned to them might partially overlap. Intercorrelation among attributes is one factor that influences clustering in an information browsing space [Dubin, 1995], and VIBE is well-suited to revealing such clustering. Whenever more than one POI has a significant influence on a document, the final position of the document icon in VIBE will be the result of a "tug of war" among the POIs. For this reason, large clusters of document icons plotted in the same region of the display suggest a dependency among two or more POIs. An analyst can reposition the POIs to confirm that the clustering is not simply an artifact of the projection onto a two-dimensional display. In Boolean VIBE, since the document icons represent distinct Boolean combinations of POIs, the problem of misinterpreting document icons that are close together in the display is minimized, if not eliminated.

Identifying correlated and independent sets of attributes is related to the goals of Principal Components Analysis and factor-analytic methods of reducing high-dimensional data spaces [Manly, 1986]. However, those methods use optimization techniques to derive linear combinations of attributes that account for as much total variability in the original similarity matrix as possible. In VIBE, attributes are combined by grouping their POI icons in the same region of the display. Unlike multivariate methods, this strategy cannot guarantee that a particular attribute combination is empirically optimal in any way. However, an analyst using VIBE can easily test alternative attribute combinations.

Ordinarily, a VIBE analyst is guided by his or her interpretation of the POI scores, and familiarity with the domain from which they are drawn. In this study, however, structure alone informed the classification of POIs into the several groups described below. Figures 3 through 13 illustrate the visual inferences that led to this classification.

5.3 The Analysis

The analysis was begun with the "Place all POIs" command of VIBE. This places all of the POIs in a circle on the screen, with the documents at various positions in the center of the circle (Figure 3). (The POIs 'A',...,'F' were not part of the original file, but were synthesized by combining POIs during the analysis. They have been deactivated for this figure.) While this display does not provide much information because it is so dense, we can make some observations from the outer regions of the figure. For example, we see that six documents contain only POI 20 (near the top), while no documents contain only POI 19. There is, however, one document placed between POIs 19 and 20, indicating that it contains only these two POIs. Similar observations can be made for other POIs.

VIBE allows the user to turn POIs on and off individually, thus altering the display. The experiment consisted of systematically exploring various POI combinations to see which produced displays that seemed informative. These displays were then "polished" by moving the POIs around to improve the display. For example, Figure 4 shows one set of POIs (1, 6, 7, 12, and 22) with a "net" invoked to give an easier interpretation of where the document icons lie. (In all figures, the numbers beside the circular POI icons are the POI numbers; the numbers beneath the rectangular document icons are the document counts. Lack of a count number indicates that a single document is present.)

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

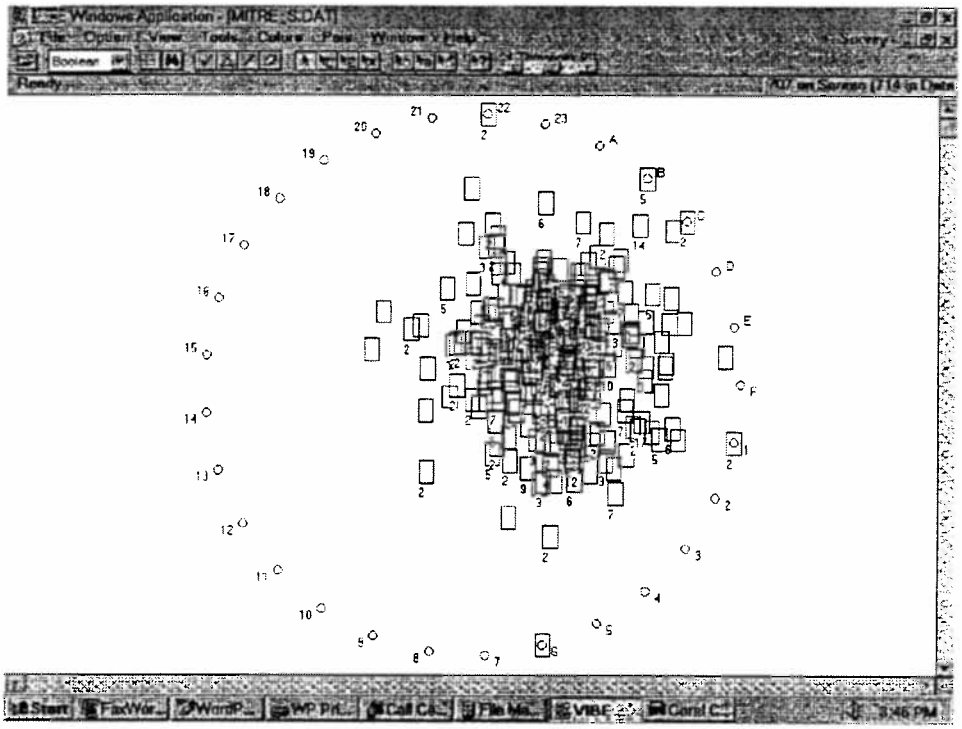

Figure 3: All POIs placed

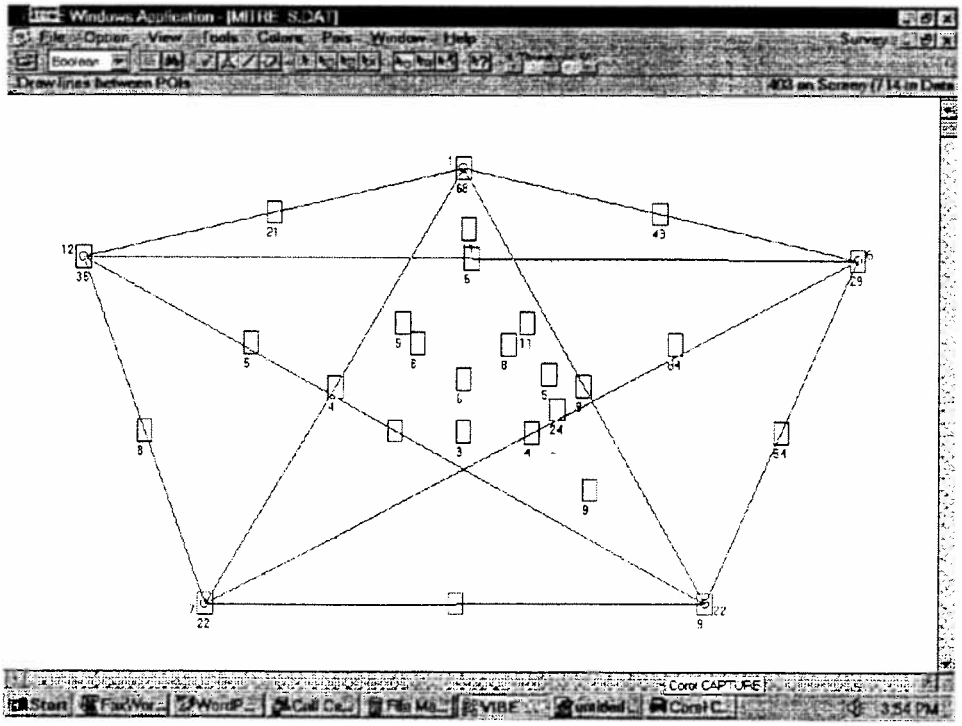

Figure 4: First POI set with "net" feature

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

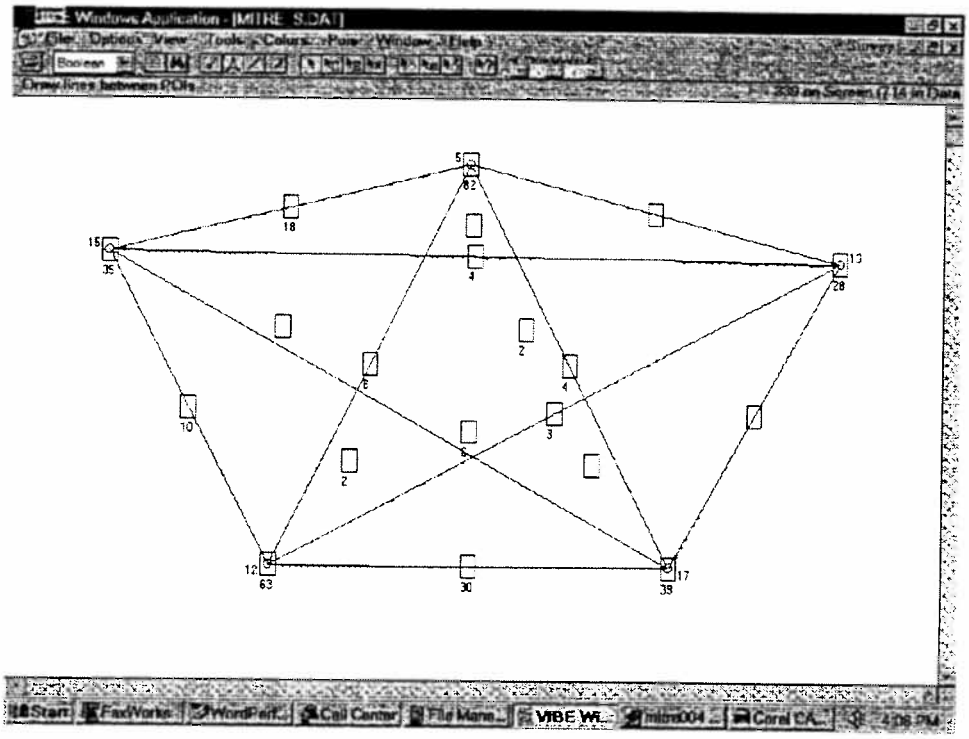

Figure 5: POI Group 2

Through this experimentation several sets of POIs were identified that seemed to display characteristics of the underlying document set structure. Following this, the results were submitted to the analyst who supplied the data for comparison with the semantics of the documentary data.

5.4 The Results

According to the VIBE-based analysis, the POIs fell largely into five groups:

- Group 1: 1, 6, 7, 12, 22
- Group 2: 5, 12, 13, 15, 17
- Group 3: 1, 6, 9, 22
- Group 4: 2, 3, 4
- Group 5: 9, 10, 16, 19, 21, 23

Figure 4 shows the first of these groups; the remaining ones are shown in Figures 5 through 11.

In Figure 4 we see that there are documents at each POI (i.e., that contain that POI and none of the other four), and documents for each pair of POIs (the midpoints of the lines). Since five POIs are used in Figure 4, there are 31 possible Boolean combinations of these POIs, not counting the null combination. This particular combination of POIs

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

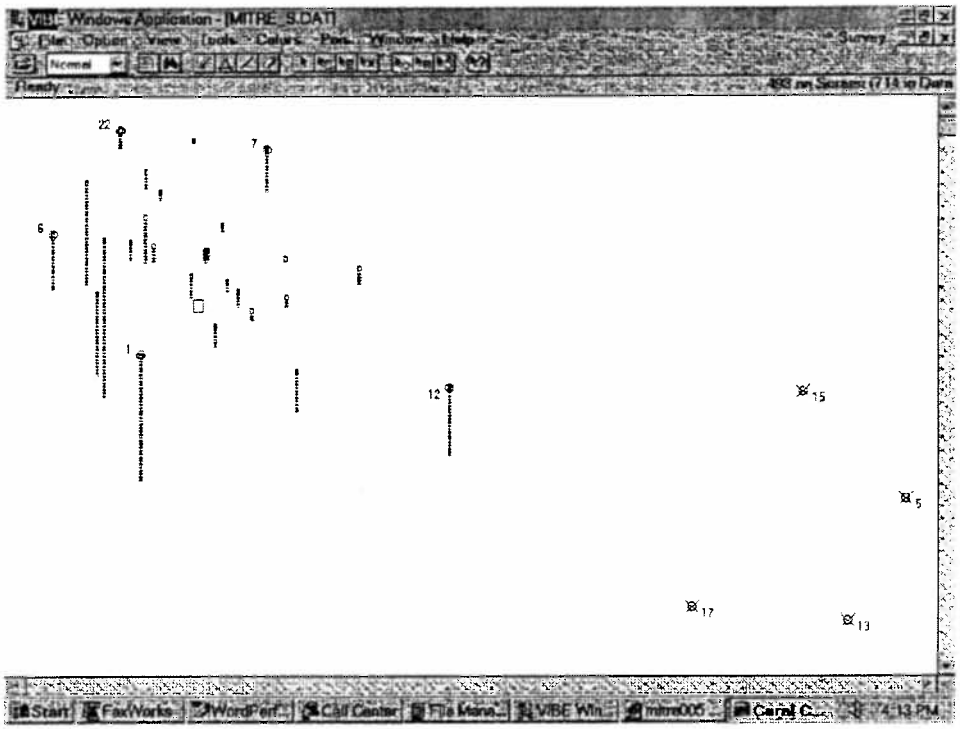

Figure 6: POI Groups 1 and 2

was chosen as possibly significant because 27 of the 31 possible combinations actually occur, thus indicating a high degree of relationship among documents with these POIs. The only combinations of POIs that do not occur are: 1-7-22, 6-7-22, 7-12-22, and 1-7-12-22. Any document containing one of these three triples also contains a fourth key term; any document containing the quadruple contains all five terms. Nearly 500 documents are represented in this figure, as shown by the numbers in the upper right corner.

Group 2, in Figure 5, is another group of documents related to five POIs, with strong internal relationships. In this case the relationships are not quite as strong, with only 20 of the possible combinations represented. This figure represents 339 documents. It should be noted that Figures 4 and 5 have one POI, 12, in common. This led to consideration of yet another configuration, consisting of Groups 1 and 2 (Figure 6). We were subsequently informed that POI 12, the shared term, is of a more general nature than the other POIs, and fits equally well with both Group 1 and Group 2.

Because of the many documents and the smaller icon size, this figure seems clearer than the corresponding Boolean representation. In Figure 7 the Group 2 POIs have been turned off, resulting in a display that is the vector equivalent of Figure 4. Comparing this to Figure 6 demonstrates the shift in document locations that results from the attraction of the Group 2 POIs. In particular, the substantial cloud of documents in the region between POI 12 and the rest of Group 1 have been pulled back into some of the larger piles of documents shown in Figure 6.

In Figure 8 one POI, 11, stands off to one side, apparently unrelated to anything else except POI 1. The analyst informed us that he had called this term "limbo," as it represented

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

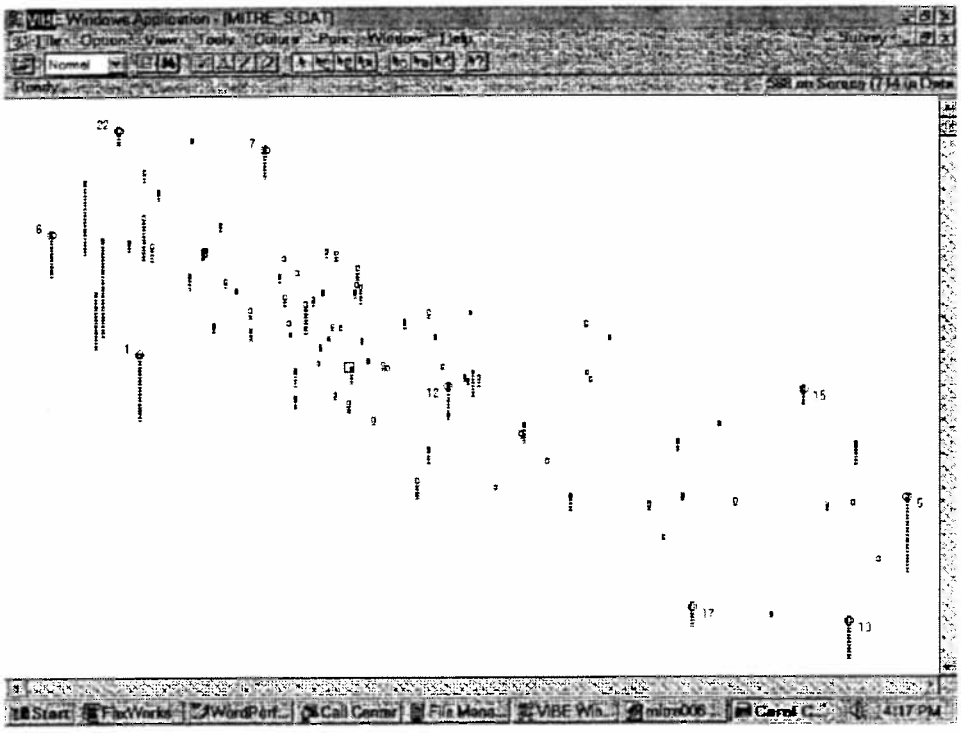

Figure 7: Group 2 POIs turned off

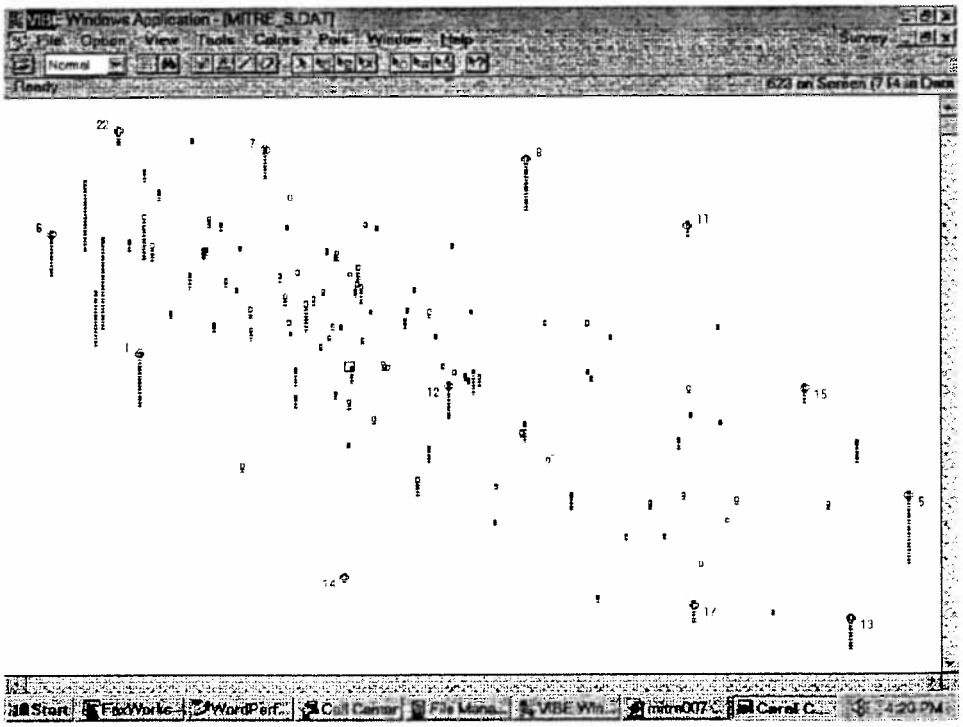

Figure 8: POI Group 3 and "limbo" POI

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

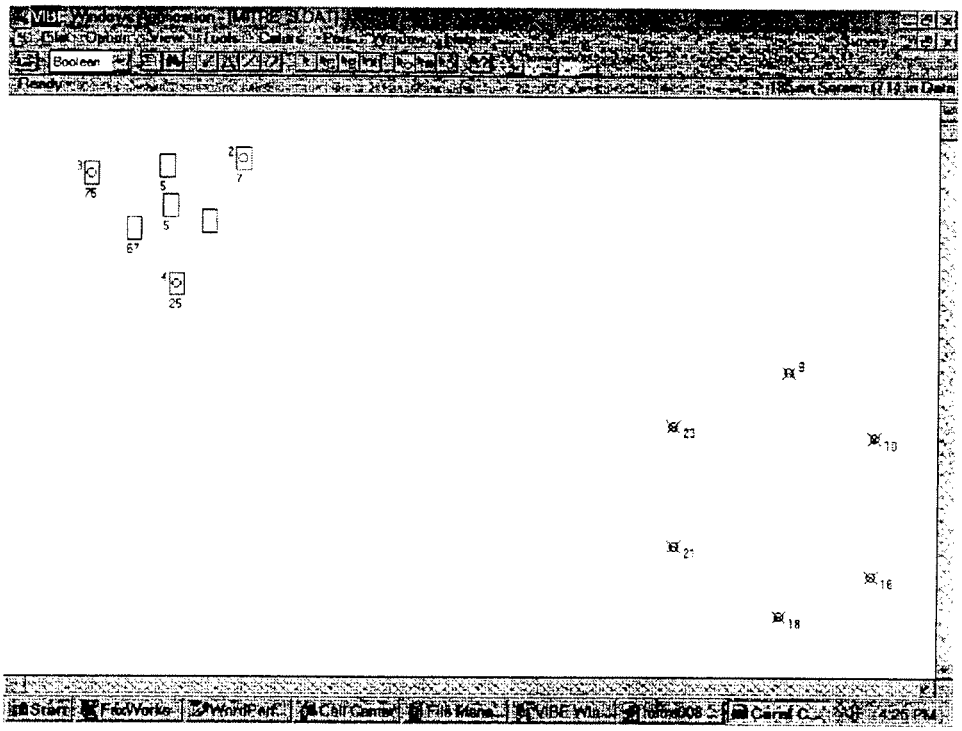

Figure 9: POI Group 4

document characteristics that did not seem to fit well with the remaining terms. This poor fit is evidenced by its visual isolation. The remaining POIs in the figure, Group 3, are very well interrelated, with 14 of a possible 15 Boolean combinations represented. (The only missing one is 9-22.) The lines in this figure are not the net joining the POIs that was used in the previous figures; rather, they are lines joining each Boolean combination to its component POIs.

Figures 9 through 11 show Groups 4 and 5 and their interaction. In Figure 9 Group 4 is shown to be fully interrelated, with all seven Boolean combinations present. Figure 10 similarly shows Group 5. In this case only 17 of a possible 63 Boolean combinations are present. The relationships are weaker, but nevertheless present. Note that in this figure POI 2 is also active; its lack of interaction with the other active POIs is evidenced by the lack of documents in the gap between the two groups. For Figure 11, POIs 3 and 4 have been activated, with the result that documents have been pulled from each of the two groups into the gap between them, disturbing the cohesion within the groups.

Figures 12 and 13 each contain Group 5 and POI 2. In Figure 12, POI 19 is deactivated, resulting in an orderly display of 279 documents. When POI 19 is activated (Figure 13) an additional 78 documents are shown (located at POI 19), and relationships among many of the original documents are modified by the influence of this new POI. However, the 18 documents associated with POI 2 remain aloof from the others.

When the analyst examined the results, he determined that each of these five groups consists of semantically related terms. The 2-3-4 group is particularly interesting since on the one hand it is tightly coupled, while on the other hand POIs 3 and 4 relate well to Group

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

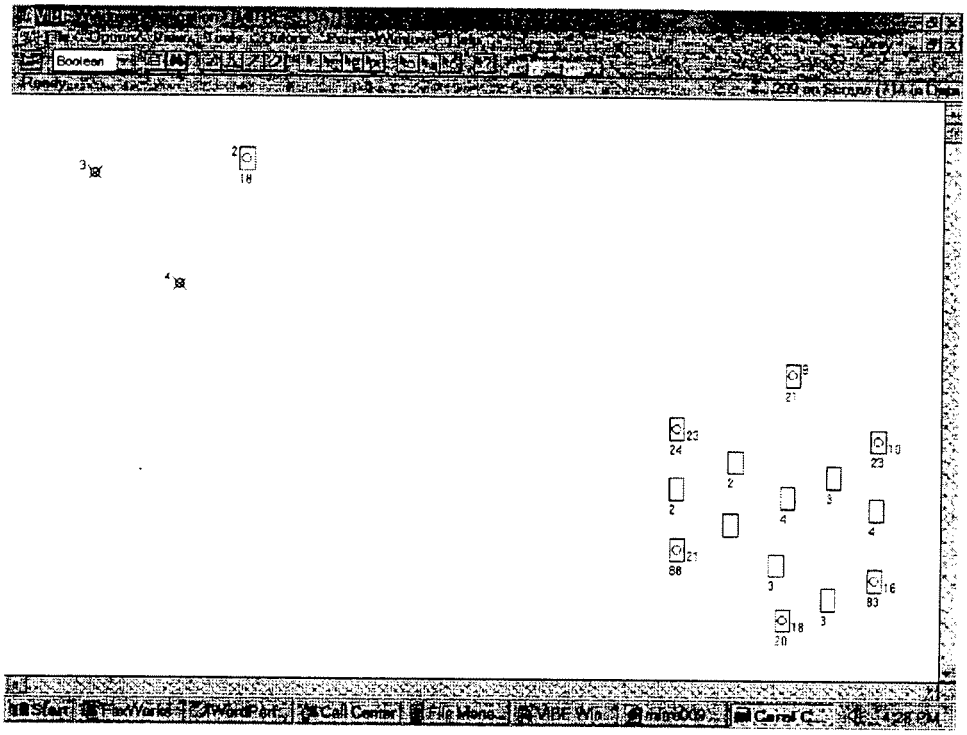

Figure 10: POI Group 5

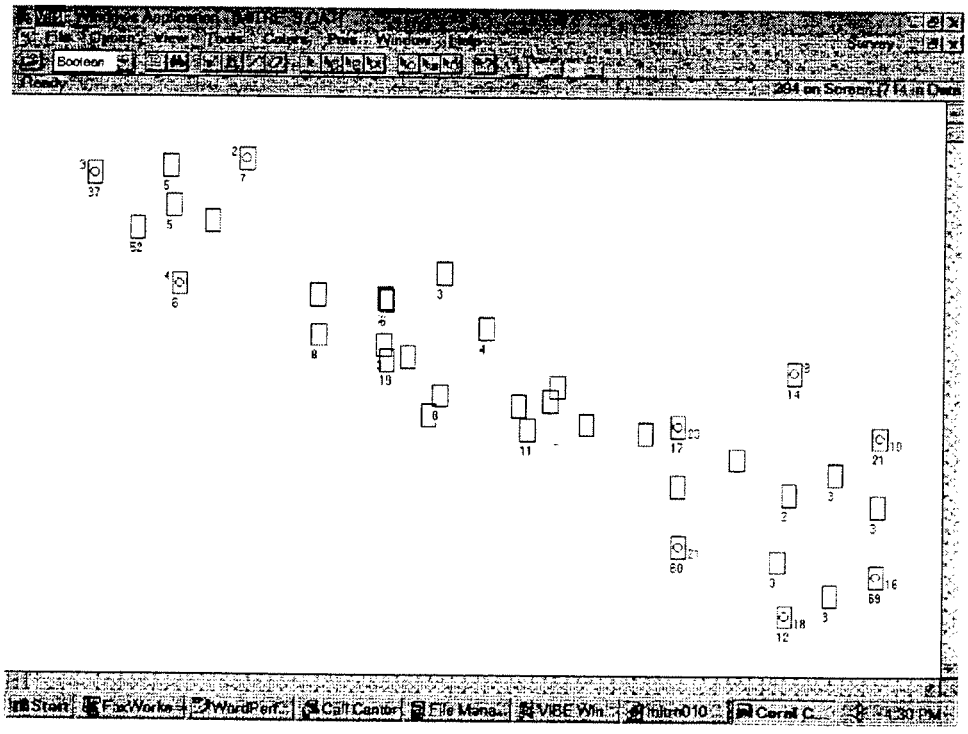

Figure 11: Effects of activating POIs 3 and 4

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

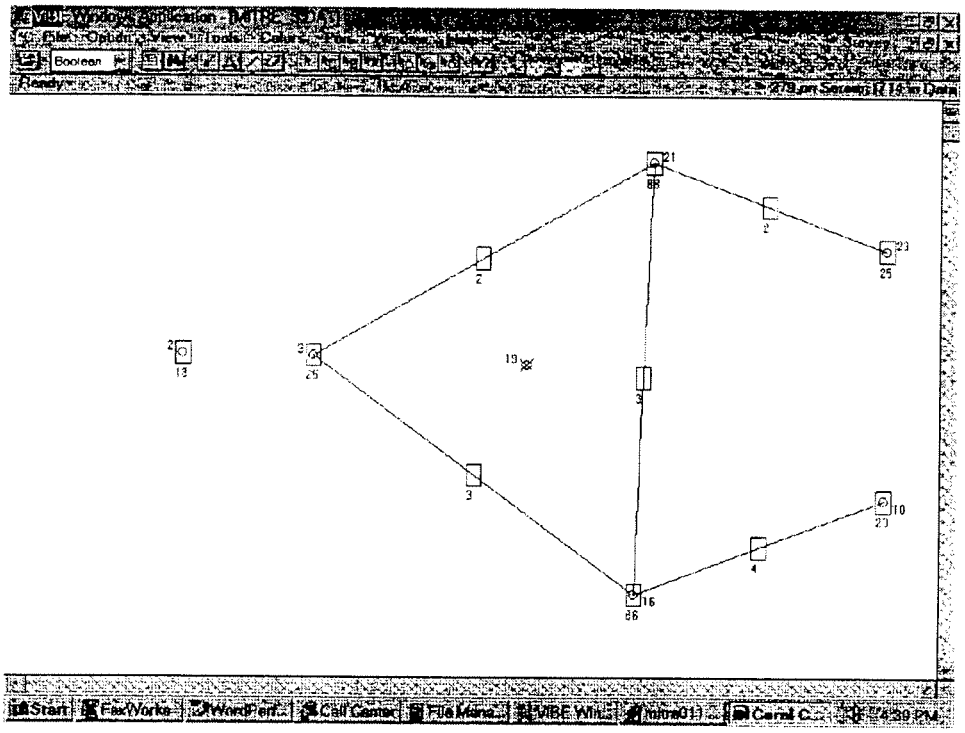

Figure 12: Group 5 and POI 2 (POI 19 non-active)

5, while POI 2 does not. It turns out that POIs 3 and 4 have a "control" aspect that is lacking from POI 2; it is apparently this aspect that causes the distinction.

Finally, if we consider the five groups together, we observe that Group 4, 2-3-4, stands by itself, but that the other groups overlap to some degree. Group 3, despite its strong internal relationships, is the most tightly bound to other groups, as three of its POIs are also in Group 1 and the fourth is in Group 5. Each of the other groups has at least one POI distinct to it.

Five of the original POIs, 8, 11, 14, 18, and 20, do not show up in any of the groups that were identified. We have already mentioned POI 11 as the "limbo" POI representing ill-fitting document characteristics. The remaining four relate to concepts of people, time, structures, and names that do not fit well with any of the concepts behind the POIs that are in the groups.

6 Conclusions

When one attempts to organize or cluster a document set on the basis of known key terms there is always the danger that the user's semantic interpretation of these terms may be influencing the perceived organization. In this study we were able to work with a set whose semantics were unknown to the experimenters, thus minimizing the potential for such bias. Nevertheless, we succeeded in identifying an organization of the documents and reference points that corresponds closely to the semantics of the document set.

Exploring the structure of the data set required a variety of POI combinations, since no

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

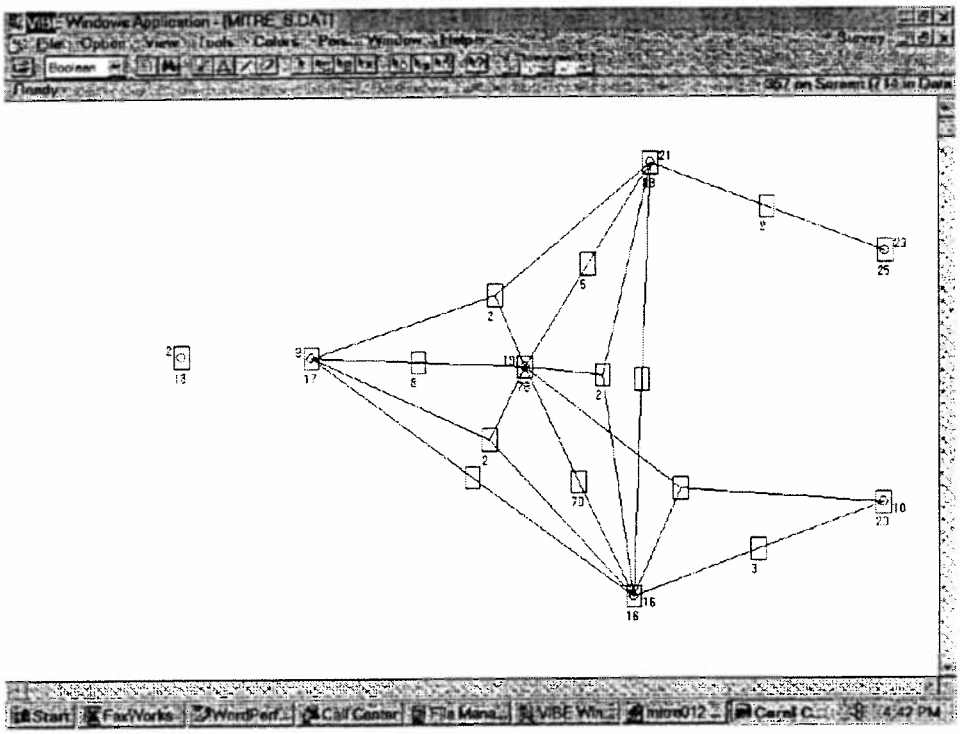

Figure 13: Group 5 and POI 2 (POI 19 active)

one configuration of POIs presents an undistorted view of the data. Although the names associated with the records and reference points were hidden, the number of reference points was small enough that combinations could be tested manually: in applications where hundreds of document attributes might define a reference point, some kind of automated filter is necessary to help separate discriminating dimensions from those contributing only noise.

The POI grouping strategy employed with the MITRE data might eventually be partially or completely automated. Clearly more experimentation must be done before this can be achieved. In particular, since our analysis technique was based on finding POI subsets with tightly coupled document sets, an automated procedure would depend on defining a measure of the coupling. We are tempted to suggest that some high percentage of the possible Boolean combinations, perhaps 85nevertheless attracted our interest.

The grouping of the POI categories in this experiment and the relationships among them discovered by use of VIBE not only conformed well to the MITRE data organization concepts, but also suggested the creation of higher level groupings, e.g. POIs 'A,'...'F,' and alternate POI designations. This work resulted in initiation of a proposal for joint research by the University of Pittsburgh and MITRE.

References

[1] Chalmers, M. & Chitson, P. (1992). Bead: Explorations in information visualization. In *Proceedings of the International ACM SIGIR Conference on Research and Development*

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

in Information Retrieval (pp. 330–337).: Association for Computing Machinery.

[2] Dubes, R. & Jain, A. K. (1979). Validity studies in clustering methodologies. *Pattern Recognition*, 11, 235–254.

[3] Dubin, D. (1995). Document analysis for visualization. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 199–204).: ACM SIGIR Association for Computing Machinery.

[4] Dubin, D. (1996). *Structure in Document Browsing Spaces*. PhD thesis, University of Pittsburgh.

[5] Lin, X. (1992). Visualization for the document space. In *Visualization 92 Proceedings* (pp. 274–281).: IEEE Computer Society Press.

[6] Manly, B. F. J. (1986). *Multivariate Statistical Methods: a Primer*. Chapman and Hall.

[7] Milligan, G. W. (1985). An algorithm for generating artificial test clusters. *Psychometrika*, 50(1), 123–127.

[8] Olsen, K. A. & Korfhage, R. R. (1994). Desktop visualization. In *Proceedings of the 1994 IEEE Symposium on Visual Languages* (pp. 239–244).: IEEE Computer Society Press.

[9] Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., & Williams, J. (1993). Visualization of a document collection: The VIBE system. *Information Processing and Management*, 29(1), 69–81.

[10] Salton, G., Yang, C. S., & Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26, 33–44.

ISSN: 2324-9773

Korfhage, R., & Dubin, D. (1996). Computer-aided Interactive Classification: Applications of VIBE. 7th ASIS SIG/CR Classification Research Workshop, 83-102. doi:10.7152/acro.v7i1.12691

ISSN: 2324-9773