COMPSTAT'2004 Symposium

ITERATIVE DENOISING FOR CROSS-CORPUS DISCOVERY

Carey E. Priebe, David J. Marchette, Youngser Park, Edward J. Wegman, Jeffrey L. Solka, Diego A. Socolinsky, Damianos Karakos, Ken W. Church, Roland Guglielmi, Ronald R. Coifman, Dekang Lin, Dennis M. Healy, Marc Q. Jacobs, Anna Tsao

Key words: Text document processing, statistical pattern recognition, dimensionality reduction.

COMPSTAT 2004 section: Dimensional reduction, Classification.

Abstract: We consider the problem of statistical pattern recognition in a heterogeneous, high-dimensional setting. In particular, we consider the search for meaningful cross-category associations in a heterogeneous text document corpus. Our approach involves "iterative denoising" — that is, iteratively extracting (corpus-dependent) features and partitioning the document collection into sub-corpora. We present an anecdote wherein this methodology discovers a meaningful cross-category association in a heterogeneous collection of scientific documents.

1 Introduction

The "integrated sensing and processing decision trees" introduced in [9] proceed according to the following philosophy. Assume that there is a heterogeneous collection of entities $\mathcal{X} = x_1, \cdots, x_n$ which can, in principle, be measured (sensed) in a large number of ways. Because the sensor cannot make all measurements simultaneously — either due to physical sensor constraints or because of the high intrinsic dimension of the complete feature collection — only a subset of the possible measurements is to be made at any one time.

Thus, for the entire entity collection $\mathcal{X}$ a first set of measurements is made. Based on the features obtained, $\mathcal{X}$ is partitioned into $\{\mathcal{X}_1, \cdots, \mathcal{X}_{J_1}\}$, each $\mathcal{X}_{j_1}$ being (presumably) more homogeneous than the original entity collection $\mathcal{X}$. Then, for each partition cell $\mathcal{X}_{j_1}$ a new set of measurements is considered. This process continues, generating branches consisting of "iteratively denoised" entity collections $\{\mathcal{X}_{j_1 1}, \cdots, \mathcal{X}_{j_1 J_2}\}$, $\{\mathcal{X}_{j_1 j_2 1}, \cdots, \mathcal{X}_{j_1 j_2 J_3}\}$, and so forth, until a collection (say, $\mathcal{X}_{j_1 j_2 j_3}$) is deemed sufficiently coherent for inference to proceed. Such collections are the leaves of the tree.

2 Iterative denoising for cross-corpus discovery

The example application we consider herein is that of discovering meaningful associations in a heterogeneous text document corpus. See, for example, [1] for a survey of text mining.

2.1 Feature extraction & dimensionality reduction

Let C be a collection of text documents. The corpus-dependent feature extraction of Lin & Pantel [6], [8] can be described as

$$\mathcal{L}_C(\cdot) : \mathcal{D}\text{ocument}\mathcal{S}\text{pace} \to [\mathcal{M}\text{utual}\mathcal{I}\text{nformation}\mathcal{F}\text{eature}]^{d_{\mathcal{L}}(C)}.$$

Both the features themselves and the number of features $d_{\mathcal{L}}(C)$ depend on the corpus C. Thus $\mathcal{L}_C(C)$ is a $|C| \times d_{\mathcal{L}}(C)$ *mutual information feature matrix*. Each of the features is associated with a word (after stemming and removal of stopper words), as follows. For document x in corpus C, and associated word w, the mutual information between x and w is given by

$$m_{x,w} = \log\left(\frac{f_{x,w}}{\sum_\xi f_{\xi,w} \sum_\omega f_{x,\omega}}\right).$$

Here $f_{x,w} = c_{x,w}/N$ where $c_{x,w}$ is the number of times word w appears in document x and N is the total number of words in the corpus C. This information is discounted to reduce the impact of infrequent words via

$$\tilde{m}_{x,w} = m_{x,w} \cdot \frac{c_{x,w}}{1 + c_{x,w}} \cdot \frac{\min(\sum_\xi c_{\xi,w}, \sum_\omega c_{x,\omega})}{1 + \min(\sum_\xi c_{\xi,w}, \sum_\omega c_{x,\omega})}.$$

The *mutual information feature vector*, then, for document x in corpus C, is given by

$$e_x = \mathcal{L}_C(x) = [\tilde{m}_{x,w_1}, \cdots, \tilde{m}_{x,w_{d_{\mathcal{L}}(C)}}].$$

Given two documents $x, y \in C$, the distance (we use the term loosely; it is in fact a pseudo-dissimilarity) employed, ρ, is given by

$$\rho(x,y) = 1 - (e_x \cdot e_y)/(||e_x||_2 ||e_y||_2) \in [0,2].$$

Thus

$$\rho \circ \mathcal{L}_C(C)$$

is a $|C| \times |C|$ *interpoint distance matrix*. All subsequent processing will be based on these interpoint distances, as discussed in [7]. However, the features, and hence the interpoint distances themselves, are *corpus dependent* and so, as the iterative denoising tree is built, based on the evolving partitioning, these distances change.

Multidimensional scaling [2] is used to embed the interpoint distance matrix $\rho \circ \mathcal{L}_C(C)$ into a Euclidean space $\mathbb{R}^{d_{mds}(C)}$. Notice first that, if the feature

vectors were Euclidean — that is, if we were using an actual distance in the $d_{\mathcal{L}}(C)$-dimensional space — then the features could be represented *with no distortion* in $\mathbb{R}^{d_{\mathcal{L}}(C)-1}$. Alas, they are not, and cannot be. So

$$mds \circ \rho \circ \mathcal{L}_C(C)$$

is a $|C| \times d_{mds}(C)$ *Euclidean feature matrix* representing the corpus C. The choice of $d_{mds}(C)$ represents a distortion/dimensionality tradeoff.

Finally, the Euclidean representation $mds \circ \rho \circ \mathcal{L}_C(C)$ produced by multidimensional scaling is reduced, via principal component analysis [5], to a lower dimensional space for subsequent processing. Again we face a model selection choice of dimensionality. The combination feature extraction/dimensionality reduction we propose, then, is given by

$$pca \circ mds \circ \rho \circ \mathcal{L}_C(C),$$

yielding a $|C| \times d_{pca}(C)$ *LSI feature matrix* which can be seen as akin to a (generalized) latent semantic indexing (LSI) [4].

2.2 Science news corpus

A heterogeneous corpus of text documents obtained from the Science News web site is used in this example. The Science News (SN) corpus C consists of $|C| = 1047$ documents in eight classes. Table 1 provides a breakdown of the corpus by number of documents per class. Our goal is two find two documents in different classes which have a meaningful association.

Class	**Number of Documents**
Anthropology	54
Astronomy	121
Behavioral Sciences	72
Earth Sciences	137
Life Sciences	205
Math & CS	60
Medicine	280
Physics	118

Table 1: Science News corpus.

For this Science News corpus C, feature extraction via $\mathcal{L}_C(C)$ yields a feature dimension $d_{\mathcal{L}}(C) = 10906$. That is, there are 10906 distinct meaningful words in the corpus, and the Lin & Pantel feature extraction produces a 1047×10906 feature matrix.

Multidimensional scaling (Figure 1, left panel) on the 1047×1047 interpoint distance matrix $\rho \circ \mathcal{L}_C(C)$ yields $d_{mds}(C) = 898$. (Numerical issues in the multidimensional scaling algorithm make 898 the largest dimension into which the interpoint distance matrix can be embedded. So, while Figure 1

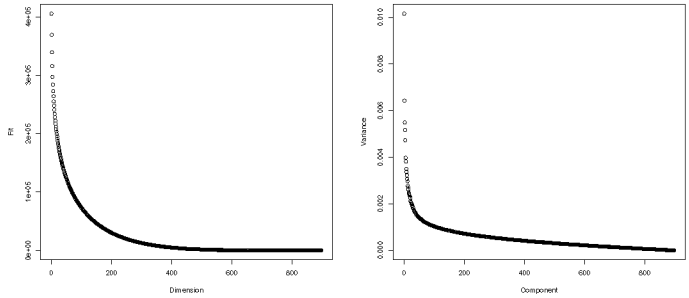

Figure 1: Multidimensional scaling (left panel) for the original 1047 10906-dimensional SN feature vectors. The largest numerically stable multidimensional scaling embedding is $d_{mds}(C) = 898$. (This left curve suggests that perhaps 200, and certainly 400 dimensions is sufficient to adequately fit the documents into Euclidean space.) Principal components (right panel) for the 898-dimensional Euclidean embedding of the original 1047 10906-dimensional SN feature vectors. (The "elbow" of this scree plot occurs, perhaps, in the range of 10-50 principal components.)

suggests that perhaps 200, and certainly 400 dimensions is sufficient to adequately fit the documents into Euclidean space, we avoid the first model selection quandary by choosing the largest numerically stable multidimensional scaling embedding.)

A subsequent principal component analysis of the 898-dimensional Euclidean features $mds \circ \rho \circ \mathcal{L}_C(C)$ yields the scree plot presented in Figure 1, right panel. This scree plot suggests that a latent semantic index dimension of perhaps 10-50 is appropriate for the SN corpus.

Figure 2 displays the projection of the data set onto the first two principal components of

$$pca \circ mds \circ \rho \circ \mathcal{L}_C(C) \tag{1}$$

for the Science News corpus. Notice that this plot suggests that the combination feature extraction/dimensionality reduction we have employed (eq. 1) has captured well some of the information concerning the eight classes, despite the fact that we are viewing just two dimensions (as opposed to, say, the 10-50 dimensions suggested by the scree plot in Figure 1). To wit: there are two groups extending from and distinguishable from the main body of documents. These two groups are dominated by medicine (the upper left arm) and astronomy (the upper right arm). Additionally, some physics documents are present in the astronomy arm and some life sciences and behavioral sciences documents are present in the medicine arm. That physics should have some similarity with astronomy, and that life sciences and behavioral sciences should have some similarity with medicine, agrees with intuition.

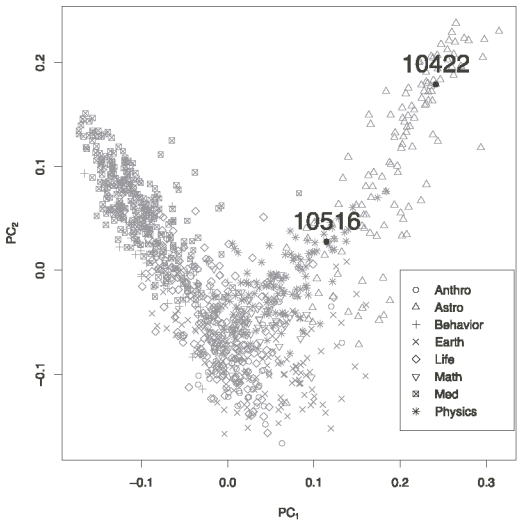

Figure 2: The first two principal components of $pca \circ mds \circ \rho \circ \mathcal{L}_C(C)$ for the Science News corpus. The eight symbols represent the eight classes; the three clusters generated via hierarchical clustering correspond roughly to the main body and the two arms. Notice that there are two groups extending from and distinguishable from the main body of documents. These two groups are dominated by medicine (the upper left arm) and astronomy (the upper right arm). The documents selected as our anecdotal "meaningful association" are indicated throughout by the solid dots and document number.

2.3 Example result

Recall that the SN corpus C has $|C| = 1047$ with class label vector

$$v = [54, 121, 72, 137, 205, 60, 280, 118].$$

The iterative denoising tree for cross-corpus discovery is illustrated on the SN corpus in Figure 3. This figure provides a coarse depiction of one path, from root to leaf, of the tree; a row-by-row description thereof follows.

Row 1: At the root, we have

$$pca \circ mds \circ \rho \circ \mathcal{L}_C(C).$$

Recall that these 1047 documents yield a feature dimension $d_{\mathcal{L}}(C) = 10906$ and an mds dimension $d_{mds}(C) = 898$. We display the first two principal components; thus the root (row 1) in Figure 3 is presented in detail in Figure 2.

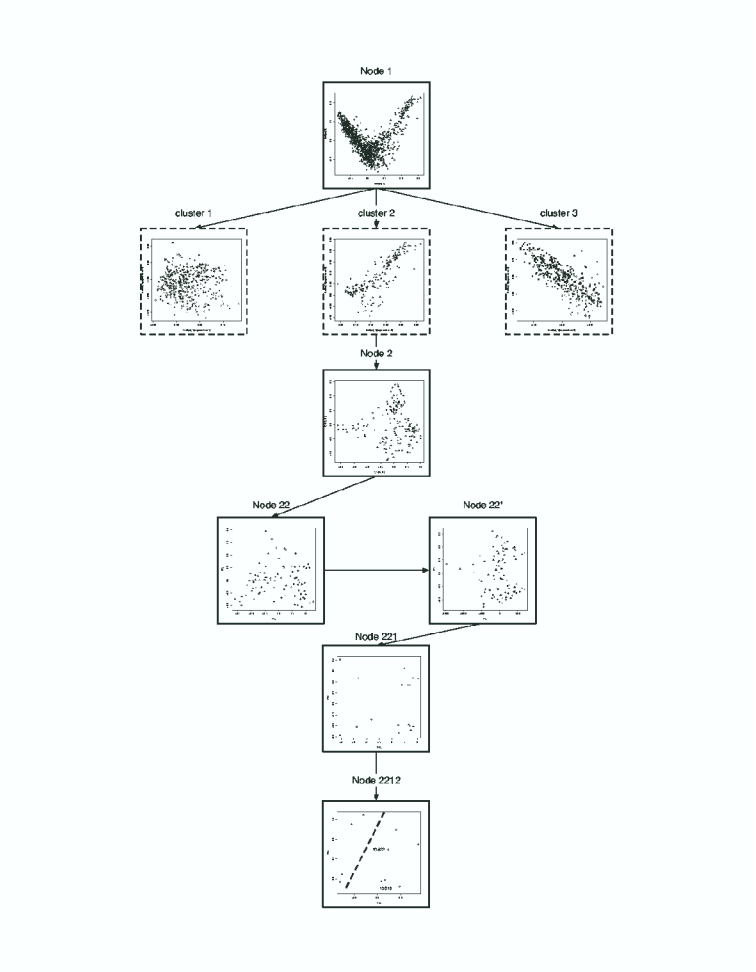

Figure 3: One path in an iterative denoising tree for the SN corpus.

Row 2: In the same space as for Row 1, we have simply split out three clusters obtained via hierarchical clustering, for display convenience.

(We choose in this manuscript to avoid model selection details; e.g., the choice of three vs. two clusters at the root. In general, we recommend that this issue be avoided by generating a *binary* tree unless user intervention is possible. In this example, the root begs for three clusters — a core and two arms.)

To illustrate an anecdotal meaningful cross-corpus discovery, we will follow cluster 2, C_2, which contains 166 documents. This subset of the original corpus is *denoised* in the sense that it is primarily physics and astronomy. The class label vector is

$$v_2 = [2, 113, 0, 10, 4, 0, 1, 36].$$

Thus, C_2 contains nearly all (113 of 121) of the astronomy documents, nearly one third (36 of 118) of the physics documents, and but a smattering from the other classes. So while the original feature extraction was done in the context of a corpus containing medicine, behavioral sciences, and mathematics documents, these topics are not a part of the context for the feature extraction for C_2 and this feature extraction can therefore focus on features germane to physics and astronomy.

Row 3: Here we display

$$pca \circ mds \circ \rho \circ \mathcal{L}_{C_2}(C_2).$$

(See Figure 4 for more detail.) These 166 documents yield a feature dimension $d_{\mathcal{L}}(C_2) = 3037$ and an mds dimension $d_{mds}(C_2) = 162$. Since $\mathcal{L}$ involves *corpus-dependent* feature extraction, this display is different than the "cluster 2" display in Row 2. This difference is due to denoising. The indicated partition represents the clusters generated via hierarchical clustering. Notice that one of the clusters (C_{22}, lower right, containing 91 documents) contains approximately half of C_2's astronomy documents (52 of 113) and nearly all of C_2's physics documents (35 of 36). In continuing pursuit of our anecdotal meaningful cross-corpus discovery, we follow C_{22}.

Row 4: The class label vector for C_{22} is

$$v_{22} = [0, 52, 0, 1, 2, 0, 1, 35].$$

The left display in Row 4 (see Figure 5 for more detail) depicts

$$pca \circ mds \circ \rho \circ \mathcal{L}_{C_{22}}(C_{22}).$$

These 91 documents yield a feature dimension $d_{\mathcal{L}}(C_{22}) = 1981$ and an mds dimension $d_{mds}(C_{22}) = 89$. Again, recall that the feature extraction is corpus-dependent. Now consider altering the geometry via the document subset

$$S_{22} = \{10500, 10651\} \subset C_{22}.$$

(These documents were chosen arbitrarily, for the purposes of illustration: they consist of a Physics document about neutrinos and an Astronomy document about black holes.) In the display, the two black squares represent S_{22}.

The right display in Row 4 (see Figure 6 for more detail) depicts the altered geometry after consideration of S_{22}. That is, here we have added

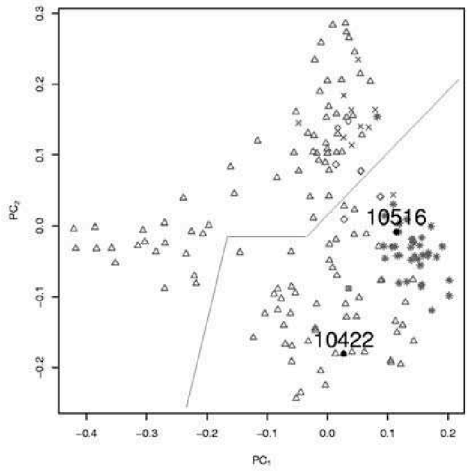

Figure 4: Node N_2 in the iterative denoising tree for the SN corpus.

a new (90th) feature $K_c d(\cdot, S_{22})$ to the 89 multidimensional scaling features, and are displaying

$$pca \circ [(mds \circ \rho \circ \mathcal{L}_{C_{22}}(C_{22}))\,; K_c d(\cdot, S_{22})]\,.$$

In the display, the two black squares again represent S_{22}. The distance-to-subset used for the additional "tunnelling" feature (see, for instance, [3]) $d(\cdot, S_{22})$, is the minimum Euclidean distance to an element of the subset in the LSI-space defined by the selected principal components; in this case, the scree plot suggests $d_{pca}(C'_{22}) = 20$. The coefficient K_c used for the tunnelling feature is obtained by scaling the values $d(\cdot, S_{22})$ so that the variance for the tunnelling feature $K_c d(\cdot, S_{22})$ is some pre-specified positive multiple c of the maximum multidimensional scaling feature variance. We use $c = 10000$ in this example so that this new feature dominates the multidimensional scaling features in the subsequent principal component analysis. (Note that the scale presented in N'_{22} in Figure 6 is such that the ordinate has no impact on the subsequent clustering; the abscissa dominates.) Rather than use the automatic clustering (depicted), we illustrate user intervention via manual clustering based on a vertical line (recall that the abscissa dominates) at 700 in N'_{22}. We follow the rightmost cluster obtained thusly, C_{221}.

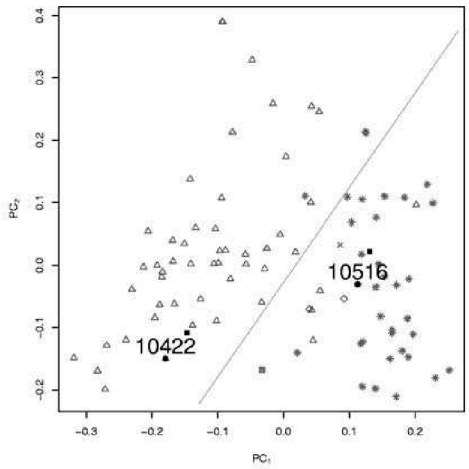

Figure 5: Node N_{22} in the iterative denoising tree for the SN corpus.

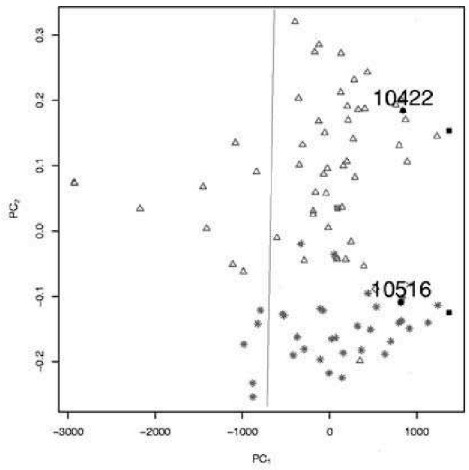

Figure 6: Node N'_{22} in the iterative denoising tree for the SN corpus.

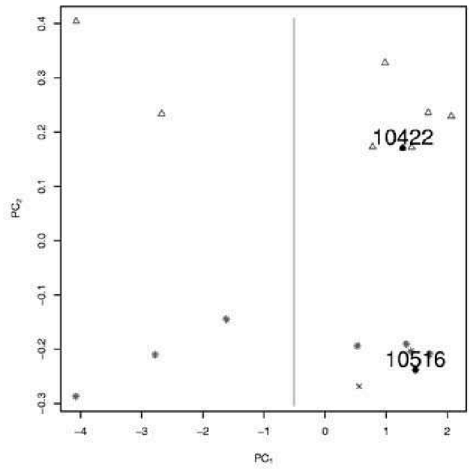

Figure 7: Node N_{221} in the iterative denoising tree for the SN corpus.

Row 5: The document collection C_{221} is, again, almost entirely astronomy and physics, with

$$|C_{221}| = 17$$

and

$$v_{221} = [0, 8, 0, 1, 0, 0, 0, 8].$$

These 17 documents yield a feature dimension $d_{\mathcal{L}}(C_{221}) = 367$ and an mds dimension $d_{mds} = 16$. After recalculating the features for C_{221}, we display

$$pca \circ [(mds \circ \rho \circ \mathcal{L}_{C_{221}}(C_{221})) ; K_{c'} d(\cdot, S_{22})] .$$

(See Figure 7 for more detail.) (A value of $c' = 100$ is used here; the impact of the tunnelling feature is lessened.)

Row 6: Here we consider one of the two clusters, C_{2212}, from N_{221} via

$$pca \circ [(mds \circ \rho \circ \mathcal{L}_{C_{2212}}(C_{2212})) ; K_{c'} d(\cdot, S_{22})] .$$

$$|C_{2212}| = 12$$

and

$$v_{2212} = [0, 6, 0, 1, 0, 0, 0, 5].$$

These 12 documents yield a feature dimension $d_{\mathcal{L}}(C_{2212}) = 215$ and an mds dimension $d_{mds} = 11$. This, in turn, clusters into C_{22121} and C_{22122}.

Let us finally consider C_{22121}. This leaf contains eight documents, with class label vector

$$v_{22121} = [0, 4, 0, 0, 0, 0, 0, 4].$$

Pairs of documents from different classes which fall to the same leaf of the iterative denoising tree are candidate associations. Thus this example yields 16 candidate associations, at least one of which (astronomy #10422 = "X-Ray Universe: Quasar's jet goes the distance" by R. Cowen, Science News Online, Feb. 16, 2002 & physics #10516 = "Glimpses inside a tiny, flashing bubble" by I. Peterson, Science News Online, Oct. 5, 1996) is plausibly a *meaningful* association.

3 Conclusion

We have presented an anecdote — not an experiment! — suggesting that an iterative denoising methodology can be a useful tool in discovering meaningful cross-corpus associations. Corpus-dependent feature extraction is an essential part of the methodology, providing features which are iteratively fine-tuned to ever more homogeneous subsets of documents as one progresses down the tree. The specific approaches to feature extraction, dimensionality reduction, and partitioning may be profitably altered within the framework of the general methodology. The adaptive geometry provided by employing distance-to-subset "tunnelling" features allows the user to alter the details of tree growth. Experimental design to allow for statistical evaluation of the performance of the methodology provides some interesting hurdles, and will be reported elsewhere.

Finally, we note that the methodology described is not specific to text document processing, and may have application in many disparate discovery scenarios. The fundamental idea, as in [9], is to address the problem of there being more measurements that can be made than should be made at any one time.

References

[1] Berry M.W., editor (2004). *Survey of text mining: clustering, classification, and retrieval*. Springer-Verlag.

[2] Borg I., Groenen P. (1997). *Modern multidimensional scaling: theory and applications*. Springer-Verlag.

[3] Cowen L.J., Priebe C.E. (1997). *Randomized nonlinear projections uncover high–dimensional structure*. Advances in Applied Mathematics **9**, 319 – 331.

[4] Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. (1990). *Indexing by latent semantic analysis*. Journal of the American Society for Information Science **41** (6), 391 – 407.

[5] Jolliffe I.T. (1986). *Principal component analysis.* Springer-Verlag.

[6] Lin D., Pantel P. (2002). *Concept discovery from text.* In Proceedings of Conference on Computational Linguistics 2002, Taipei, Taiwan, 577 – 583.

[7] Maa J.-F., Pearl D.K., Bartoszynsky R. (1996). *Reducing multidimensional two-sample data to one-dimensional interpoint comparisons.* The Annals of Statistics **24**, 1069 – 1074.

[8] Pantel P., Lin D. (2002). *Discovering word senses from text.* In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002, Edmonton, Canada, 613 – 619.

[9] Priebe C.E., Marchette D.J., Healy D.M. (2004). *Integrated sensing and processing decision trees.* IEEE Trans. PAMI, to appear.

Acknowledgement: Sponsored by the Defense Advanced Research Projects Agency under "Novel Mathematical and Computational Approaches to Exploitation of Massive, Non-physical Data", ARPA Order No. P246, Program Code 3E20. Issued by DARPA/CMO under Contract No. MDA972-03-C-0014 to AlgoTek, Inc. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either explicitly or implied, of DARPA or the U.S. Government. Approved for Public Release, Distribution Unlimited.

Address: C.E. Priebe, E.J. Wegman, D.A. Socolinsky, K.W. Church, R. Guglielmi, R.R. Coifman, D. Lin, M.Q. Jacobs, A. Tsao, AlgoTek, Inc., 3811 N. Fairfax Dr., Suite 700
D.J. Marchette, J.L. Solka, NSWCDD B10, Dahlgren, VA
Y. Park, D. Karakos, Johns Hopkins U., Balt., MD
D.M. Healy, DARPA, Arlington, VA 22203

E-mail: cep@jhu.edu