A Text Stream Transformation for Semantic-Based Clustering

Angel R. Martinez
George Mason University[1]

Edward J. Wegman
George Mason University

Abstract:

Assuming a bounded domain of discourse, a transformation of the text stream, called a bigram proximity matrix, is proposed. The proximity matrix is used to classify documents using k nearest neighbor (kNN) discrimination, and various distances are evaluated. To use other types of classifiers, we reduce the dimensionality of the proximity matrix using the Isomap method of nonlinear dimensionality reduction. Classification results on the reduced data are also presented. From the results presented in this paper, it seems that the bigram proximity matrix preserves meaning.

Key Words: *k* nearest neighbor classification, model-based clustering, text retrieval, representation of meaning, natural language processing, change of topic

1. Introduction

This paper introduces a non-textual representation of the text stream in a high-dimensional construct called the bigram proximity matrix (BPM). Computational attributes of the construct allow for efficient manipulation by computational statistics methods. Possible applications of the approaches described here include: text retrieval and topic change determination.

We first provide some background information on previous work in natural language research. This is followed by a description and example of the BPM. Subsequent sections describe experiments conducted to determine whether the BPM is a structure that preserves meaning. These experiments involve *k*NN classification, clustering, and hypothesis testing. Results from these experiments are presented and conclusions are offered.

2. Background

Approximately 35 years ago, the quest to endow computers with the capacity to understand natural language began. The effort was originally called Natural Language Understanding (NLU). Now, it is most frequently called Natural Language Processing (NLP). The change reflects a lowering of expectations from the original goals of the AI research community.

At the beginning of these efforts, the introduction of Chomsky's theory of generative grammars opened the door to the possibility of automatically generating all possible utterances of English. This provided the impetus for the following approach to automatic knowledge acquisition:

1. Parse the natural language utterance,
2. Manipulate its syntactic constituents, and
3. Group meaning-carrying components into formal knowledge representation structures.

This approach makes two basic assumptions: (1) a grammar of English can be defined and automatically used to parse any utterance in the language, and (2) contextual knowledge can be stored and used for automated elucidation

1. Email: martinezar@nswc.navy.mil (Angel Martinez) and ewegman@galaxy.gmu.edu (Edward Wegman)

of meaning. Eventually, it became evident that the richness of the language disallowed the possibility of complete, unambiguous, and computational efficient methods for capturing knowledge through the simple rules of generative grammars [Révézs, 1983]. Similar results have plagued the work of inducing machine understanding by creating a vast semantic context [Landauer, Laham, and Foltz, 1998], [Charniak, 1996].

The unfulfilled expectations of the original AI-NLP work motivated some researchers to look beyond the classic computational semantics approaches of the linguists. As Charniak states, it was "therefore time to switch paradigms." The paradigm shift took the form of a statistical approach to NLP. Two main, sometimes overlapping, threads of statistical NLP are evident now: (1) the capture and representation of semantics, and (2) information retrieval.

This paper introduces a meaning-representation transformation of the text stream into structures amenable to computational tools. Experiments in supervised and unsupervised learning will demonstrate the extent to which the structures capture meaning. Although application of the structure in text retrieval and topic change determination seem possible, the hope is that these structures, or like structures, can be used in the development of NLP methods where semantic features, more than syntactic features, are used in the processing of natural languages.

3. The Bigram Proximity Matrix

The bigram proximity matrix (BPM) is a non-symmetric matrix that captures the number of words' co-occurrences in a moving 2-word window. The BPM, an innovation of this work, is a square matrix whose column and row headings are the alphabetically ordered entries of the lexicon. Each matrix element is the number of times word *i* appears immediately before word *j* in the unit of text. The size of the BPM is determined by the size of the lexicon created by listing alphabetically the unique occurrences of the words in the text. It is our assertion that the BPM representation of the semantic content preserves enough unique features to be semantically separable from BPMs of other thematically unrelated collections.

We must make some comments about the lexicon and the pre-processing of the documents before proceeding with more information on the BPM. All punctuation within a sentence such as commas, semi-colons, colons, etc., are removed. All end-of-sentence punctuation, other than a period, such as question marks and exclamation points are converted to a period. The period is used in the lexicon as a 'word,' and it is placed at the beginning of the alphabetized lexicon.

Table 1. Example of Bigram Proximity Matrix[a]

	.	crowd	his	in	father	man	sought	the	wise	young
.										
crowd	1									
his					1					
in								1		
father				1						
man							1			
sought			1							
the		1							1	
wise										1
young						1				

a. Zeros in empty boxes are removed for clarity.

The BPM is a simple structure that seems to preserve much of the semantics of the originating text. Recall from the beginning of this section that the rows in the BPM represent the first word in the pair, and the second word is shown in the column. For example, the BPM for the sentence or text stream,

"The wise young man sought his father in the crowd."

is shown in Table 1. We see that the matrix element located in the third row (*his*) and the fifth column (*father*) has a value of one. This means that the pair of words *his father* occurs once in this unit of text. It should be noted that in most cases, depending on the size of the lexicon and the size of the text stream, the BPM will be very sparse.

By preserving the ordering of words of the discourse stream, the BPM captures a substantial amount of information about meaning. Also, by obtaining the individual counts of word co-occurrences, the BPM captures the 'intensity' of the discourse's theme. Both features make the BPM a suitable tool for capturing meaning and performing computations to identify semantic similarities among units of discourse (e.g., paragraphs, documents). Determining how well the BPM captures meaning is one of the main objectives of this research.

4. Description of Experiments

In order to determine if the BPM captures meaning, three sets of experiments where conducted.

- Set 1: Supervised learning experiments using the *k*NN classification method
- Set 2: Unsupervised learning experiments with model-based clustering
- Set 3: Formal hypothesis tests via a graph-theoretic approach and Monte Carlo replications

Documents from the Topic Detection and Tracking (TDT) Pilot Corpus (Linguistic Data Consortium, Philadelphia, PA) were used as the textual testbed. The TDT corpus is comprised of close to 16,000 stories collected from July 1, 1994 to June 30, 1995 from the Reuters newswire service and CNN broadcast news transcripts. A set of 25 events are defined in the TDT. Each of the 16,000 stories is flagged with one of three possible flags: *Yes*, *No*, or *Brief*. The flags are used to indicate that a story discusses one of the 25 events, or it does not, or it does so only briefly. These topic tags were obtained by having two people read the stories and classify them accordingly.

In order to meet computational requirements, a subset of the TDT corpus was used in this work. A total of 503 stories were chosen from the 16,000 available. These stories comprise 16 of the 25 events discussed in the TDT. See Table 2 for a list of topics. The 503 stories chosen contain only the *Yes* or *No* flags. This choice stems from the need to demonstrate that the BPM captures enough meaning to make a correct or incorrect topic classification choice.

The 503 stories selected produced a lexicon of 11,103 unique words. Conflated forms are counted as different words. Using this lexicon and the structure already described, a BPM was created for each of the 503 stories. The assertion is that each of the BPMs captures enough of the meaning in the story to serve as a classification feature. The rest of this dissertation research is devoted to testing this assertion using various methods in exploratory data analysis and computational statistics.

Table 3 lists the measures of semantic similarity used in this study. For definitions of these measures, see [Martinez, 2002]. It should be noted that some of these are distances and some are similarities; however, for ease of exposition, we will refer to them all as measures of semantic similarity. All similarity measures were first converted to distances so they could be used in the classification and dimensionality reduction methods. Also, some of these measures are binary, in which case the bigram frequency is changed to a 0 or a 1. Similarly, the BPMs are converted to distributions to use the probabilistic measures.

Table 2. List of 16 Topics

Topic Number	Topic Description	Number of Documents Used	Topic Number	Topic Description	Number of Documents Used
4	Cessna on the White house	14	15	Kobe, Japan Quake	50
5	Clinic Murders (Salvi)	38	16	Lost in Iraq	30
6	Comet into Jupiter	45	17	NYC Subway bombing	24
8	Death of Kim Jong Il	35	18	Oklahoma City bombing	76
9	DNA in OJ trial	29	21	Serbians down F-16	16
11	Hall's copter in N. Korea	75	22	Serbs violate Bihac	19
12	Humble, TX, flooding	16	24	US Air 427 crash	16
13	Justice-to-be Breyer	8	25	WTC bombing trial	12

Table 3. Measures of Semantic Similarity

1. Matching Coefficient	8. Sokal - Sneath
2. Sokal - Michner	9. Gower - Legendre 1
3. Dice Coefficient	10. Gower - Legendre 2
4. Jaccard Coefficient	11. Normalized Correlation Coefficient
5. Cosine - Ochiai	12. L_1 - Probabilistic Measure
6. Russell - Rao	13. IRad - Probabilistic Measure
7. Roger - Tanimoto	

Two variants of the BPM are considered. In one variant, common high-frequency words have been removed from the lexicon and the documents. In another variation, we stemmed the words as well as removed the common high-frequency words from the documents. Part of this research examines how this affects the discriminating power of the BPMs.

Many NLP applications [Kimbrell, 1988], [Salton, Buckley and Smith, 1990], [Frakes and Baeza-Yates, 1992], [Berry and Browne, 1999] use a shorter version of the lexicon by excluding words often used in the language. These words, usually called *stop words* or *noise words*, are said to have low informational content and thus, in the name of computational efficiency, are deleted. Not all agree with this approach [Witten, Moffat and Bell, 1994].

In order to determine how the semantic representation power of the BPM are affected by eliminating noise words from the lexicon, new proximity matrices were computed and distances regenerated using a denoised lexicon. The expectation is that by eliminating words with high frequencies and low semantic content, the rarer words would increase the discriminatory factor of the features.

Taking the denoising idea one step further, we stemmed the words in the denoised text and then created the corresponding BPMs. The idea is to conflate words to their stem or root to increase the frequency of key words and thus enhance the discriminatory factors of the features. Stemming is routinely applied in the area of information retrieval (IR). In the IR application, stemming is used to enhance the performance of the IR system, as well as to reduce the total number of unique words and save on computational resources.

A popular stemmer is the Porter stemmer [Baeza-Yates and Neto, 1999], [Porter, 1980]. The Porter stemmer is simple; however, its performance is comparable with older established stemmers. Our implementation of the Porter stemmer in MATLAB proved to be more than adequate and very fast. The Porter stemmer works on the suffix of words. These are stripped according to several parsing rules and replaced with one of a list of endings or the null ending. On most occasions, these simple replacements work well, as is the case of the words: *protecting, protected, protects, protection.* These words are easily conflated to the stem '*protect.*' However, a word like *probate* will be stemmed to *probe*, which carries a totally different meaning. The same happens for *relativity,* which is conflated to *relate*. In IR, these anomalies of the stemmers seem not to detract from their use. The issue is not so clear in our application. To determine if there is any deleterious effect of using the stemmer and denoising with the BPMs, the same experiments are conducted on all three versions: full, denoised, and stemmed documents. Table 4 summarizes the size of the lexicons in these three cases

It is important to note the different purpose for stemming in our work compared to that in IR systems. Although both applications are related to identifying documents by their word composition, our work focuses on the 'identification' of meaning as embodied in documents. The distinction is important because the results of the Porter stemmer (or *any* stemmer) can work against our goal while at the same time support the IR goals.
.

Table 4. Lexicon Sizes

Type of Lexicon	Size of Lexicon
Full Lexicon	11,103
Denoised Lexicon	10,997
Stemmed Lexicon (also denoised)	7,146

5. Results from Experiments

In this section, we describe the three sets of experiments outlined in the previous section. Results are also discussed. All experiments were implemented in MATLAB in a PC-Windows environment.

5.1 *k*NN Classification

We first apply supervised learning approaches to the BPM features to determine whether these features allow us to classify documents according to their meaning. If we can accurately classify using the proximity matrix as a feature, then this is an indication that these features preserve meaning.

Due to the high-dimensional nature of these features, we cannot apply classical supervised learning methods such as linear or quadratic classifiers. However, the *k* nearest neighbor (*k*NN) classifier is well-suited for these features since all we need are the pairwise distances between all BPMs.

The decision rule is to assign $\mathbf{x}$ to class ω_m if

$$\hat{p}(\omega_m|\mathbf{x}) \geq \hat{p}(\omega_i|\mathbf{x}), \tag{1}$$

for all *i*. Or, using Bayes' Theorem and our probability estimates, we can restate the rule as

$$\frac{k_m}{n_m \mathrm{V}} \frac{n_m}{n} \geq \frac{k_i}{n_i \mathrm{V}} \frac{n_i}{n}, \tag{2}$$

for all *i*, where k_m is the number of samples that belong to ω_m, n_m is the total number of samples in ω_m, n is the total number of samples, **V** is the volume centered at **x** determined by the fixed value of *k*, and *C* is the total number of classes. Thus, **x** is assigned to *m* if

$$k_m \geq k_i, \tag{3}$$

for all *i*. In other words, the decision rule is to assign **x** to the class that has the greater number of members amongst the *k* nearest neighbors [Web, 1999], [Cover and Hart, 1967].

The *k*-nearest neighbor classification method was applied using the following:

Text Condition:	full, denoised, and stemmed
k Values:	1, 3, 5, 7, and 10
Proximity Matrix Type:	BPM
Measures of semantic similarity:	all as listed in Table 3

The values shown in Tables 5 to 7 are the ratio of the number of correct classification to the number of possible classifications. From now on, this metric, the correct classification ratio, will be abbreviated as CCR. Of the thirteen measures of semantic similarity, 10 resulted in high CCRs. Three of them resulted in very low CCRs. These are the Sokal-Michner, Roger-Tanimoto, and Gower-Legendre 1.

Summary conclusions from the above observations are:

1. The BPM seems to contain sufficient semantic information to allow for almost perfect classification results.
2. Binary similarity measures based solely on the logical operation **X** AND **Y** worked better with denoised text
3. Probabilistic measures of semantic similarity performed the best with denoised and stemmed text.
4. The Dice, Jaccard, Sokal-Sneath, Gower-Legendre 2, L_1 norm and IRad measures were the best performers.

Table 5. Results of *k*NN - Full Lexicon (Bold values indicate the highest overall CCR.)

Similarity Measures	$k = 1$	$k = 3$	$k = 5$	$k = 7$	$k = 10$
Matching	0.7197	0.2316	0.8429	0.8708	0.9006
Sokal-Michner	0.2942	0.2127	0.2545	0.1948	0.2306
Dice	0.9722	0.9742	0.9801	0.9761	0.9761
Jaccard	0.9722	0.9742	0.9801	0.9761	0.9761
Ochiai	0.9821	**0.9841**	**0.9841**	**0.9841**	0.9821
Russell-Rao	0.7197	0.7316	0.8429	0.8708	0.9006
Roger-Tanimoto	0.2942	0.2127	0.2545	0.1948	0.2306
Sokal-Sneath	0.9722	0.9742	0.9801	0.9761	0.9761
Gower - Legendre 1	0.1014	0.0616	0.0537	0.0557	0.0517
Gower- Legendre 2	0.9722	0.9742	0.9801	0.9761	0.9761
Normalized Corr Coefficient	0.8986	0.8986	0.9105	0.9245	0.9284
L_1 Norm	0.9642	0.9583	0.9662	0.9602	0.9642
IRad	0.9662	0.9642	0.9602	0.9642	0.9622

Table 6. Results of *k*NN - Denoised Lexicon (Bold values indicate the highest overall CCR.)

Similarity Measures	$k = 1$	$k = 3$	$k = 5$	$k = 7$	$k = 10$
Matching	0.9523	0.9563	0.9682	0.9722	0.9722
Sokal-Michner	0.3002	0.2724	0.2664	0.1690	0.2167
Dice	0.9881	0.9781	0.9881	0.9781	0.9821
Jaccard	0.9881	0.9781	0.9881	0.9781	0.9821
Ochiai	**0.9920**	0.9861	0.9861	0.9861	0.9781
Russell-Rao	0.9523	0.9563	0.9682	0.9722	0.9722
Roger-Tanimoto	0.3002	0.2724	0.2664	0.1690	0.2167
Sokal-Sneath	0.9881	0.9781	0.9881	0.9781	0.9821
Gower - Legendre 1	0.1014	0.0636	0.1014	0.0795	0.0696
Gower- Legendre 2	0.9881	0.9781	0.9881	0.9781	0.9821
Normalized Corr Coefficient	0.9284	0.9284	0.9284	0.9284	0.9185
L_1 Norm	0.9881	0.9801	0.9821	0.9761	0.9742
IRad	0.9901	0.9821	0.9841	0.9761	0.9761

Table 7. Results of *k*NN - Stemmed Lexicon (Bold values indicate the highest overall CCR.)

Similarity Measures	$k = 1$	$k = 3$	$k = 5$	$k = 7$	$k = 10$
Matching	0.9483	0.9443	0.9742	0.9722	0.9761
Sokal-Michner	0.3201	0.2724	0.2744	0.1829	0.2207
Dice	0.9841	0.9801	0.9801	0.9781	0.9881
Jaccard	0.9841	0.9801	0.9801	0.9781	0.9881
Ochiai	0.9901	0.9841	0.9861	0.9861	0.9881
Russell-Rao	0.9483	0.9443	0.9742	0.9722	0.9761
Roger-Tanimoto	0.3201	0.2724	0.2744	0.1829	0.2207
Sokal-Sneath	0.9841	0.9801	0.9801	0.9781	0.9881
Gower - Legendre 1	0.1014	0.0636	0.0557	0.0676	0.0676
Gower- Legendre 2	0.9841	0.9801	0.9801	0.9781	0.9881
Normalized Corr Coefficient	0.9443	0.9384	0.9423	0.9404	0.9304
L_1 Norm	0.9881	0.9801	0.9821	0.9801	0.9781
IRad	**0.9920**	0.9881	0.9841	0.9821	0.9781

5.2 Model-Based Clustering

The method chosen for our unsupervised learning experiments is called model-based clustering [Banfield and Raftery, 1993]. This method is based on finite mixtures [Everitt and Hand, 1981] where the output model is a weighted sum of c multivariate normals:

$$f(\mathbf{x}) = \sum_{k=1}^{c} p_k \phi(\mathbf{x}; \mu_k, \Sigma_k) . \tag{4}$$

The general idea is to generate estimates based on Equation 4, where constraints are imposed on the covariance matrices. The best estimate and model (i.e., number of components, parameter estimates, and form of the covariance matrices) is chosen based on the model that yields the highest value of the Bayesian Information Criterion (BIC), given by

$$BIC \equiv 2L_M(\mathbf{x}, \hat{\theta}) - m_M \log(n) , \tag{5}$$

where m_M is the number of parameters in model M and L_M is the log likelihood. The final model and estimate we deem the best will be the one that corresponds to the highest value of the BIC. See Figure 1 below, obtained from one of the experiments.

Before we outline the steps of model-based clustering, we return to the issue of the constraints imposed on the covariance matrices that give rise to the various models (see Table 8). Our implementation of model-based clustering used four of the possible models. These are outlined and described in Table 8.

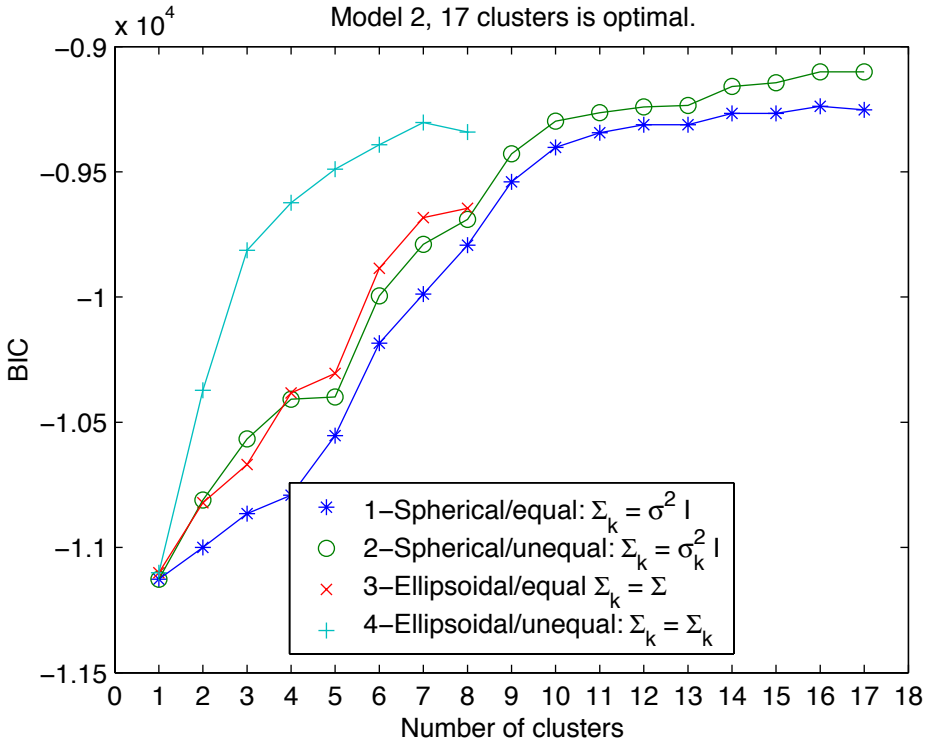

Figure 1. Here we show the BIC values for the experiment where the Jaccard measure is used, denoised text, 6-D, and k = 7 nearest neighbors (Isomap). Note that the highest BIC corresponds to 17 clusters and model 2.

Table 8. Description of the Four Models Used in This Research

Model Number (M)	Covariance	Model	Description
1	Spherical and equal	$\hat{\Sigma}_k = \sigma^2 \mathbf{I}$	• Diagonal covariance matrices • Same value in diagonal elements • Covariance matrices are equal
2	Spherical and unequal	$\hat{\Sigma}_k = \sigma_k^2 \mathbf{I}$	• Diagonal covariance matrices • Covariances are allowed to vary between components • Same value in each diagonal element of individual covariance matrix
3	Ellipsoidal and equal	$\hat{\Sigma}_k = \Sigma$	• Covariance matrices can have non-zero off-diagonal elements • Covariance matrices are equal
4	Ellipsoidal and unequal	$\hat{\Sigma}_k = \Sigma_k$	• Covariance matrices can have non-zero off-diagonal elements • Covariance matrices can vary among components

As a way to compare supervised and unsupervised learning methods with the use of BPMs, similar experiment variables used with the *k*NN classification were used with model-based clustering. The following variable combinations were used:

- Thirteen measures of semantic similarity (see Table 3)
- Two values of *k* nearest neighbors for the Isomap dimensionality reduction: $k = 7, 10$
- Proximity matrix: BPM
- Three text conditions: full, denoised and stemmed
- One 'best' dimension value from Isomap

In order to use model-based clustering, the dimensionality of our observations (i.e., the BPMs) had to be drastically reduced from $11{,}103^2$ (in the case of the full lexicon) to 2, 3, 4, 5, and 6 dimensions. This reduction was effected through the Isometric Figure Mapping (Isomap), a nonlinear dimensionality reduction method [Tenenbaum, deSilva and Langford, 2000].

Following the suggestion of Fraley and Raftery [Fraley and Raftery, 1998], the unconstrained model was used with the agglomerative clustering method to initialize the EM algorithm. The four models were then used, with values of *c* given by $c = 2, 3, 4, ..., 18$. A total number of 72 BIC values were obtained for the determination of the best model with number of components and parameter estimates.

The assessment of the results was done via a visualization aid we developed called ReClus. ReClus takes the output from the model-based clustering procedure and draws one large rectangle. This rectangle is subdivided into *n* smaller rectangles, where *n* is the number of clusters chosen according to the highest BIC value. The area of each smaller rectangle is proportional to the number of cases in the cluster. Inside each rectangle, and for each case assigned to that cluster, the class number is printed, or optionally, the case number is printed. Each number is color-coded to denote the degree of certainty that the particular case belongs to the cluster. A threshold is set to print in black bold type when the certainty is 0.8 or above.

ReClus, thus, provides a quick visual way to examine the results from model-based clustering. Although, judging between two results entails a degree of subjectivity, this is a problem only where results are close. Additionally, ReClus provides information to guide the examination of confounding factors in the clustering process.

The relevant results are listed below in Table 9. These are those where the number of clusters (or terms in the mixture) suggested as optimal by the BIC were close to 16, since this is the correct number of clusters in the test bed used for the experiments. For each experiment, the number of 'pure' clusters with cases belonging to only one class, and clusters with minimal 'contamination' from other classes, i.e., almost 'pure' clusters, were counted. Three more categories were counted in increasing degrees of 'contamination': medium, heavy, and jumbled (a thorough mix of various classes).

We now offer some specific observations on the results, keeping in mind that it is difficult to assess the goodness of clusters. Of the 312 experiments, thirteen showed the correct number of clusters, sixteen. Not surprisingly, however, none of these - and for that matter, none of the 312 - showed sixteen correct (i.e., 'pure') clusters. In each of the thirteen results, two rectangles contained the same class cases (topic number 6). We note that the same situation for topic 6 arose in those results containing 15 and 17 clusters. Usually, more than half of the rectangles suffered from some degree of contamination.

Table 9. Ratings of Relevant Model-Based Clustering Results

Measure	Text	Num of Dim[a]	k[a]	Number of Clusters	Subjective Assessment of Contamination				
					Pure	Light	Med	Heavy	Jumbled
Match	Den	4	7	14	1	4	1	1	7
Dice	Den	6	7	16	6	3	3	2	2
Jaccard	Den	6	7	17	7	4	2	1	3
Ochiai	Den	6	7	17	7	5	1	1	3
Sokal-Sneath	Den	6	7	17	7	4	2	0	4
G-L 2	Den	6	7	16	6	5	1	2	2
Dice	Stem	6	7	16	6	5	2	1	2
Jaccard	Stem	6	7	17	7	5	2	1	1
Ochiai	Stem	6	7	17	7	5	2	0	3
G-L 2	Stem	6	7	15	7	2	1	1	4
IRad	Stem	2	7	15	7	2	1	1	4
Ochiai	Full	5	10	16	6	4	1	1	4
Ochiai	Den	5	10	16	6	5	1	0	4
Ochiai	Full	6	7	17	6	6	2	2	1
Ochiai	Full	5	7	15	6	5	1	0	3

a. These are from Isomap. For the dimensions, it indicates the best dimension. For the *k,* it indicates the number of nearest neighbors used in the Isomap algorithm.

If we consider a good result as one with the highest number of ‘pure’ rectangles, followed by a high number of only lightly ‘contaminated’ ones, and the fewest number of jumbled rectangles, then the following are the best results:

- Ochiai measure, full text, dimensionality 6, BPM and $k = 7$
- Jaccard measure, stemmed text, dimensionality 6, BPM, and $k = 7$

The above categorization of the best results is naive. It assumes that a mix of 2 or more classes in a rectangle is an undesirable result. However, in the case of our test bed, a mix could point to a justifiable confusion. For example, in several of the ‘best’ results cases 8 and 11 are usually mixed; however, both sets of documents are about North Korea. Also, topics 18 and 17 are sometimes mixed: both sets of documents deal with bombing, the Oklahoma City bombing and the NY subway bombing. The same happens a few times with cases 21 and 22: both report on two different aspects of the Serbian conflict.

The intriguing case mentioned above, where class 6 had two pure rectangles containing class 6 cases, raises the issue of latent classes or sub-groups within the topics. A reading of the documents involved does show two different foci. The main subject of the set is the crash of fragments of the comet Shoemaker-Levy onto the surface of Jupiter. One group in the set emphasizes background information about the comet as well as the fact that the space shuttle is in orbit ready to observe what is yet to take place. The second group’s focus is predominantly on the event already taking place and observations of the phenomenon.

The observations discussed in the two paragraphs above, were clearly validated via parallel coordinates [Wegman, 1990]. We devised a matrix of parallel coordinates for simultaneous observation of the sixteen topics (clusters). See Figure 2 at the end of the paper. By looking at the overall shape formed by the lines and the points where these touch the five axes (5 dimensions), we are able to detect patterns. These patterns seem to be manifestations of semantic content of the clusters. Notice the following:

- The parallel coordinates for topics 8 and 11 show exact patterns for a good number of their lines. This corroborates the confusion detected in the model-based clustering results via the ReClus display (Figures 4 and 5). The possible common theme repeated is North Korea and US relations.
- The parallel coordinate plots for topics 17 and 18 show a group of lines with the exact pattern in both. This corroborates the confusion detected in the model-based clustering results via ReClus (Figures 4 and 5). A possible common theme that is repeated is bombing and its immediate effects.
- The parallel coordinates for topics 21 and 22 show a small group of lines with a common pattern. This pattern may represent a common core of the two topics about the Serbian conflict.
- Topic 6 showed invariably in two clusters in the ReClus figures (Figures 4 and 5). Notice the pattern from the lines of the parallel coordinates for topic 6. On the second axis from top to bottom, one notices a separation of lines. This indicates two different groups, separable at the dimension represented by that axis. These groups may represent the two sub-themes found in reading the newscasts of topic 6.

Parallel coordinates were a crucial help in making sense of our model-based clustering results. On occasions, a permutation tour [Wegman, 1990] of a single topic's parallel coordinates was necessary. For example, the parallel coordinate plot for topic 8 shows a simple structure. But, the ReClus view shows topic 8 linked with topic 11 in three different clusters. Is the model-based clustering result wrong? When a permutation tour of topic 8 was performed, about three sub-groups became evident as can be seen in Figure 3 at the end of the article. This seems to match a reading of the 35 newscasts from topic 8 as the next paragraph explains.

As mentioned above, classes 8 and 11 appeared mixed in the experiments. Topic 8 and topic 11 both deal with North Korea, one regarding the death of Kim Il Sung and the other the crash of the American helicopter in North Korean territory. Most of the time there are three rectangles containing cases from 8, of which two are mixed with 11 and one rectangle (almost purely 11) was only very slightly mixed with 8. As is the case with class 6, this may imply the existence of latent classes in groups 8 and 11. A quick reading of the newscasts for topic 8 seems to show three major themes discussed over the background of Kim Il Sung's death and the probable succession of his son Kim Jong-il. The three latent topics are: (1) US and North Korea relations and nuclear issues talks; (2) North Korea and South Korea relations; and (3) memorial services and mourning in North Korea and South Korea.

The visualization tool ReClus made the examination of the results from the model-based clustering experiment possible and fruitful. The BPMs capture sufficient meaning to produce satisfactory results with this unsupervised learning method. For best results, the Ochiai measure of semantic similarity should be used in the Isomap dimensionality reduction method, and the dimensionality can be reduced to five or six dimensions. Full and denoised text did well with the Ochiai measure. It seems that latent classes are detected by the BPM, as made manifest by the results discussed above. Below (Figures 4 and 5) are examples of the ReClus output from some of the experiments.

5.3 Nettleton-Bannerjee Hypothesis Test

A more rigorous test of the capacity of the proximity matrices to capture meaning was performed as a hypothesis test. The underlying concept of the test is the equality or inequality of two or more distributions of random proximity matrices (BPMs). Our assumption is that documents that belong to one topic will belong to the same distribution as measured by the proximity matrices.

Nettleton and Bannerjee [Nettleton and Bannerjee, 2001] propose a graph-theoretic method for hypothesis testing where nearest neighbors are linked by edges and where the test statistic is the number of edges linking observations from different distributions. In what follows, this method will be called the N-B method.

The N-B method, intended for binary vectors, was adapted to our BPMs. To accommodate the method, all values greater than zero were converted to one. The general problem at hand can be described as follows [Nettleton and Bannerjee, 2001].

For $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$, let $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,k})$, be an observation from distribution F_i. Thus, we have I distributions, each containing n_i vectors, each of dimensionality K. For $k = 1, \ldots, K$, suppose $\mathbf{X}_{ijk}$ takes integer values from 0 to $m_k - 1$ for some integer $m_k \geq 2$. Thus, $\mathbf{X}_{ij}$ can take any of $M \equiv \prod_{k=1}^{K} m_k$ vector values. In our case, $I = 16$, $K = 11103^2$, $m_k = 2$ for all k, and $M = 2^{11103^2}$. Referring back to Table 4, one can calculate the corresponding values for the denoised and stemmed lexicons.

Let F_i be the distribution (or topic) to which BPM_i belongs, then the hypothesis test performed was the following:

$$\text{H}_0: \quad F_i = \bigcup_{i \neq j} F_j \qquad i = 1, \ldots, 16, \; j = 1, \ldots, 16, \quad i \neq j$$

$$\text{H}_1: \quad F_i \neq \bigcup_{i \neq j} F_j \qquad i = 1, \ldots, 16, \; j = 1, \ldots, 16, \quad i \neq j$$

This hypothesis test creates two groups, one comprised of only one distribution (or topic), while the other group is comprised of the union of the remaining distributions (or topics). It then tests to see if the i-th distribution is equal to the union of the rest.

The measures of semantic similarity listed in Table 3 were used in the hypothesis test. In general, δ_k defines a distance function

$$\delta(\mathbf{X}_{ij}, \mathbf{X}_{st}) = \sum_{k=1}^{K} \delta_k(\mathbf{X}_{ijk}, \mathbf{X}_{stk}) . \tag{6}$$

These distances or similarity measures are used to determine where to draw an edge to link any two data points, that is

$$\delta(\mathbf{X}_{ij}, \mathbf{X}_{st}) = \min \{\delta(\mathbf{X}_{ij}, \mathbf{X}_{uv}): u = 1, \ldots, I; v = 1, \ldots, n_u; (u, v) \neq (i, j)\}$$

or

$$\delta(\mathbf{X}_{ij}, \mathbf{X}_{st}) = \min\{\delta(\mathbf{X}_{uv}, \mathbf{X}_{st}): u = 1, \ldots, I; \; v = 1, \ldots, n_u; (u, v) \neq (s, t)\} .$$

The test statistic T is the number of edges that cross group boundaries. Thus, if we let $D_\delta(\mathbf{X}_{ij}, \mathbf{X}_{st}) = 1$ if the points are linked by an edge and 0 otherwise, then

$$T = \sum_{i=1}^{I-1} \sum_{s=i+1}^{I} \sum_{j=1}^{n_i} \sum_{t=1}^{n_s} D_\delta(\mathbf{X}_{ij}, \mathbf{X}_{st}) . \tag{7}$$

A small value of T provides evidence against H_0. The significance of T was assessed by comparing the observed value of T to its conditional null distribution.

Monte Carlo simulation was used to approximate the p-value associated with the observed value of T. This was accomplished efficiently by randomly permuting group labels in 100,000 replications or trials. Our estimate of the p-value was obtained by calculating the proportion of the Monte Carlo replicates of the test statistic for which the observed T value is greater than or equal to the replicated value of T.

Results are shown in Table 10. These hypothesis tests induce a grouping of the similarity measures into four groups. The first group is comprised of the Dice, Jaccard, Ochiai, Sokal-Sneath, Gower-Legendre 2, normalized correlation coefficient, L_1, and IRad similarity measures. This group produced results supportive of the alternative hypothesis for each and every one of the sixteen topics. The T statistic values were concentrated in the range of $0 \le T \le 25$, with several of the measures producing the exact same values. The second grouping comprised of the match and the Russell-Rao coefficients performed almost as well as the first group, with the exception of the p-values for topics 8 and 17.

The Sokal-Michner and the Roger-Tanimoto coefficients produced almost exact results. Compared to the numbers of the previous two groups, their performance was not impressive. Values for topics 4, 13, 17, 18, 21, 22, and 25 indicate support for the null hypothesis. The Gower-Legendre 1 coefficient stands by itself as the worst performer by far. This similarity measure did not work well with the BPM to provide enough discrimination power. It managed to perform well in one case, separating topic 4 from the rest. In general, the BPMs in combination with ten of the thirteen similarity measures provided good discrimination power, indicating the BPMs capacity to preserve meaning of the text stream.

6. Summary, Conclusions, and Future Work

We introduced in this paper the bigram proximity matrix (BPM), a transformation of the text stream amenable to computational methods. The usefulness of the BPM depends on how much semantic information it preserves. In order to determine the adequacy of the BPM to preserve semantic information, three sets of experiments were performed: (1) supervised learning using kNN classification, (2) unsupervised learning, using model-based clustering, and (3) a hypothesis test. Variables in the experiments consisted of combinations of the following:

- Thirteen semantic similarity measures
- Three text conditions (full, denoised and stemmed)
- Various values of k (kNN).

Supervised learning experiments, as well as the hypothesis test, were conducted on the full dimensionality of the feature space (see Table 4 for the lexicon sizes). Dimensionality was reduced to a lower number (2 - 6) using the nonlinear dimensionality reduction procedure called Isometric Figure Mapping (Isomap). With dimensionality reduced, unsupervised learning experiments were conducted using model-based clustering.

Results from supervised learning experiments showed that correct classification ratios in the range of 0.95 - 0.99 were common for many of the semantic similarity measures used. This indicates that the BPM captures sufficient semantic information for the discrimination of semantically dissimilar text units. Results from the unsupervised learning experiments showed that the BPM captures sufficient semantic information to group thematically related documents and seems to detect latent sub-themes. The hypothesis test, used to test if the BPM preserved enough semantic information to discriminate each class member from the other classes, produced unequivocal results to reject the hypothesis that members of class i are the same (semantically) as the members of all other classes. In conclusion, we can state that the text stream transformation BPM does capture enough semantic information to allow for the semantic discrimination of text units.

Several obvious possibilities for future work are:

- To extend the BPM to a trigram proximity matrix (TPM). To do this, we can create a cube by adding a third dimension also defined by the lexicon.
- To apply BPMs and TPMs to the problem of change of topic determination.
- To explore the capability of BPMs and TPMs in combination with model-based clustering, parallel coordinates and ReClus in the detection and identification of sub-topics.

Table 10. Results for Hypothesis Test

Topic Label	Dice, Jaccard, Ochiai, S-S, G-L2, NCC, L_1, IRad		Matching Coefficient Russell-Rao (these produce essentially the same results)		Sokal-Michner Roger-Tanimoto (these produce essentially the same results)		Gower-Legendre 1 Similarity Measure	
	T	p-Value	T	p-value	T	p-value	T	p-value
4	$0 \le T \le 25$	0	0	0	21	0.86	4	0
5	$0 \le T \le 25$	0	9	0	41	.0004	50	0.7794
6	$0 \le T \le 25$	0	6	0	30	0	44	0.7595
8	$0 \le T \le 25$	0	116	0.93	16	0	435	0.9940
9	$0 \le T \le 25$	0	2	0	31	0.029	29	0.8360
11	$0 \le T \le 25$	0	26	0	78	.0007	74	0.6200
12	$0 \le T \le 25$	0	7	0	13	.0001	16	0.9060
13	$0 \le T \le 25$	0	5	0	366	0.98	8	0.9537
15	$0 \le T \le 25$	0	23	0	52	0.005	49	0.7370
16	$0 \le T \le 25$	0	2	0	32	0.024	30	0.8312
17	$0 \le T \le 25$	0	31	0.37	29	0.41	24	0.8634
18	$0 \le T \le 25$	0	46	0	126	0.74	76	0.6119
21	$0 \le T \le 25$	0	3	0	19	0.42	16	0.9085
22	$0 \le T \le 25$	0	8	0	24	0.61	19	0.8903
24	$0 \le T \le 25$	0	7	0	15	0.002	16	0.9077
25	$0 \le T \le 25$	0	3	0	13	0.19	12	0.9315

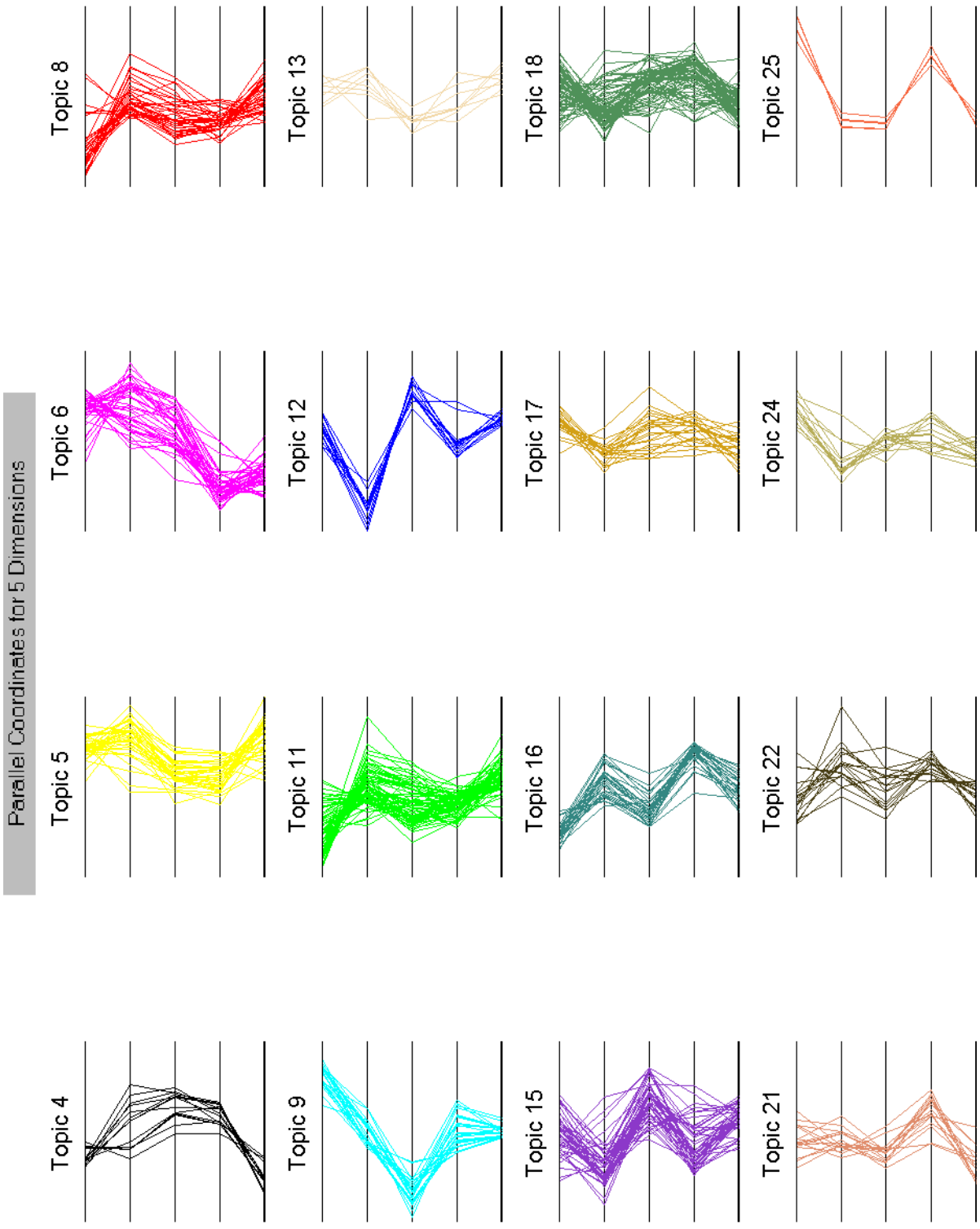

Figure 2. Parallel coordinates plot matrix for the 16 topics.

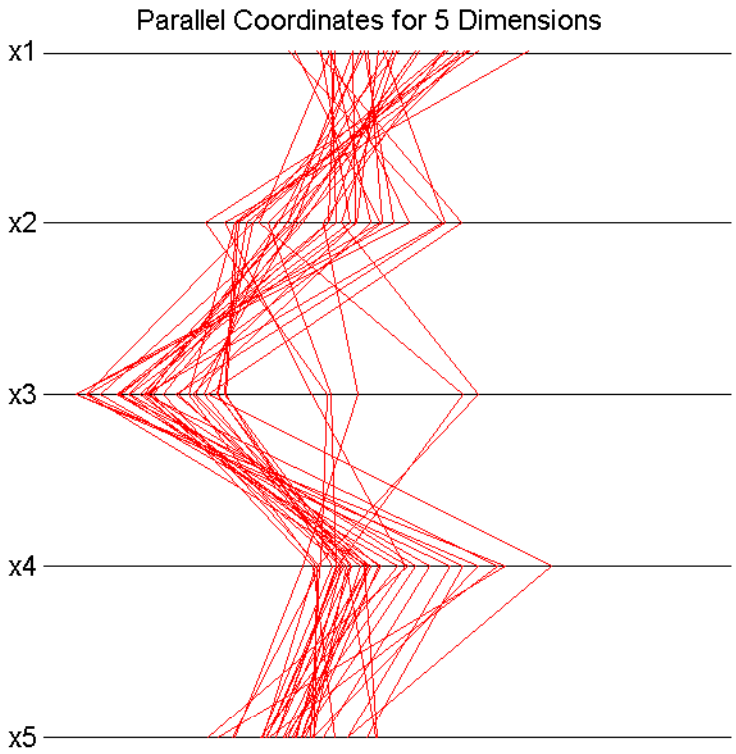

Figure 3. This shows the parallel coordinates plot for Topic 8 only, after the axes have been permuted. Notice the groups are now visible on one of the axes.

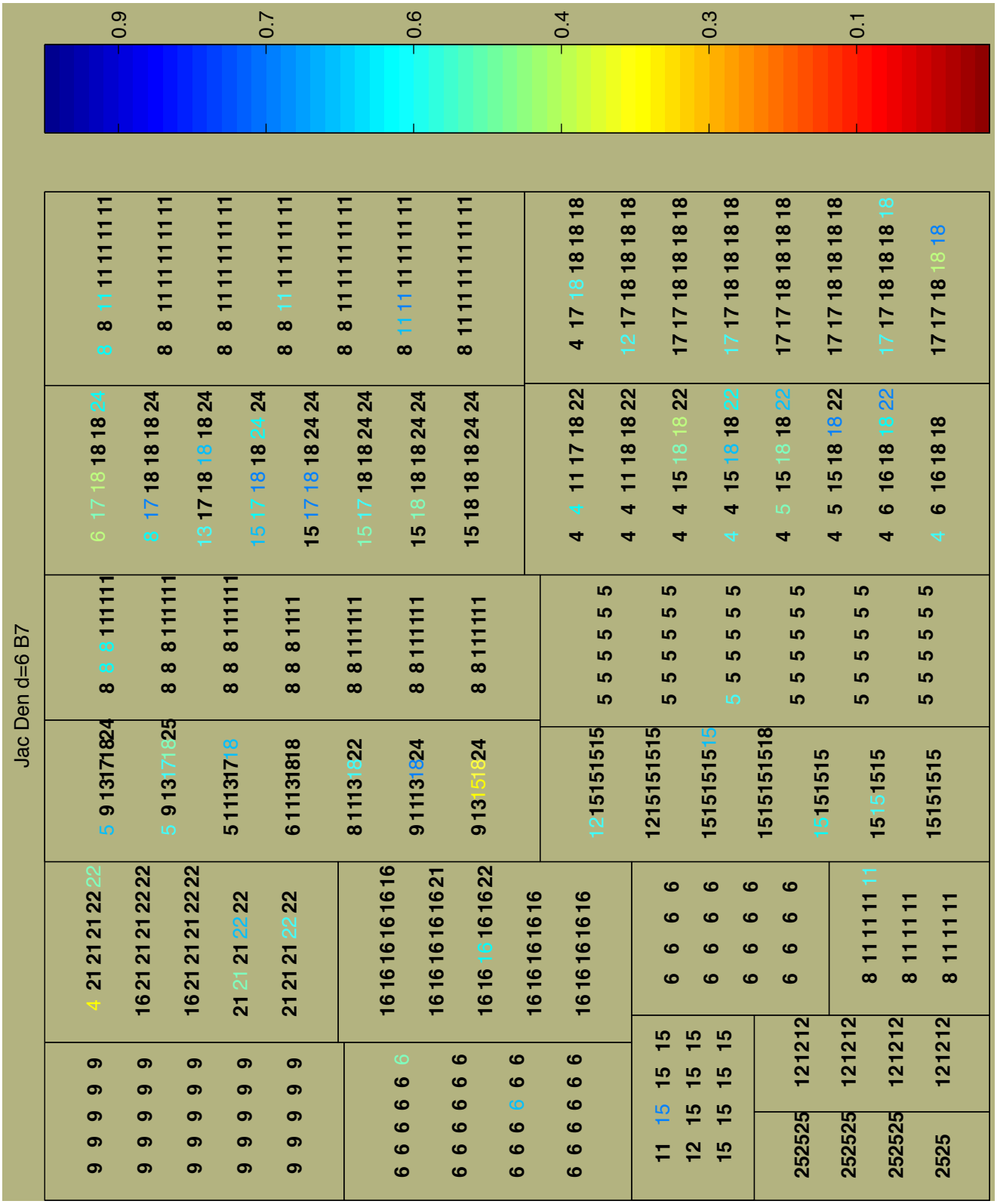

Figure 4. Here we see the ReClus layout showing the results from the model-based clustering where the Jaccard measure is used with denoised text.

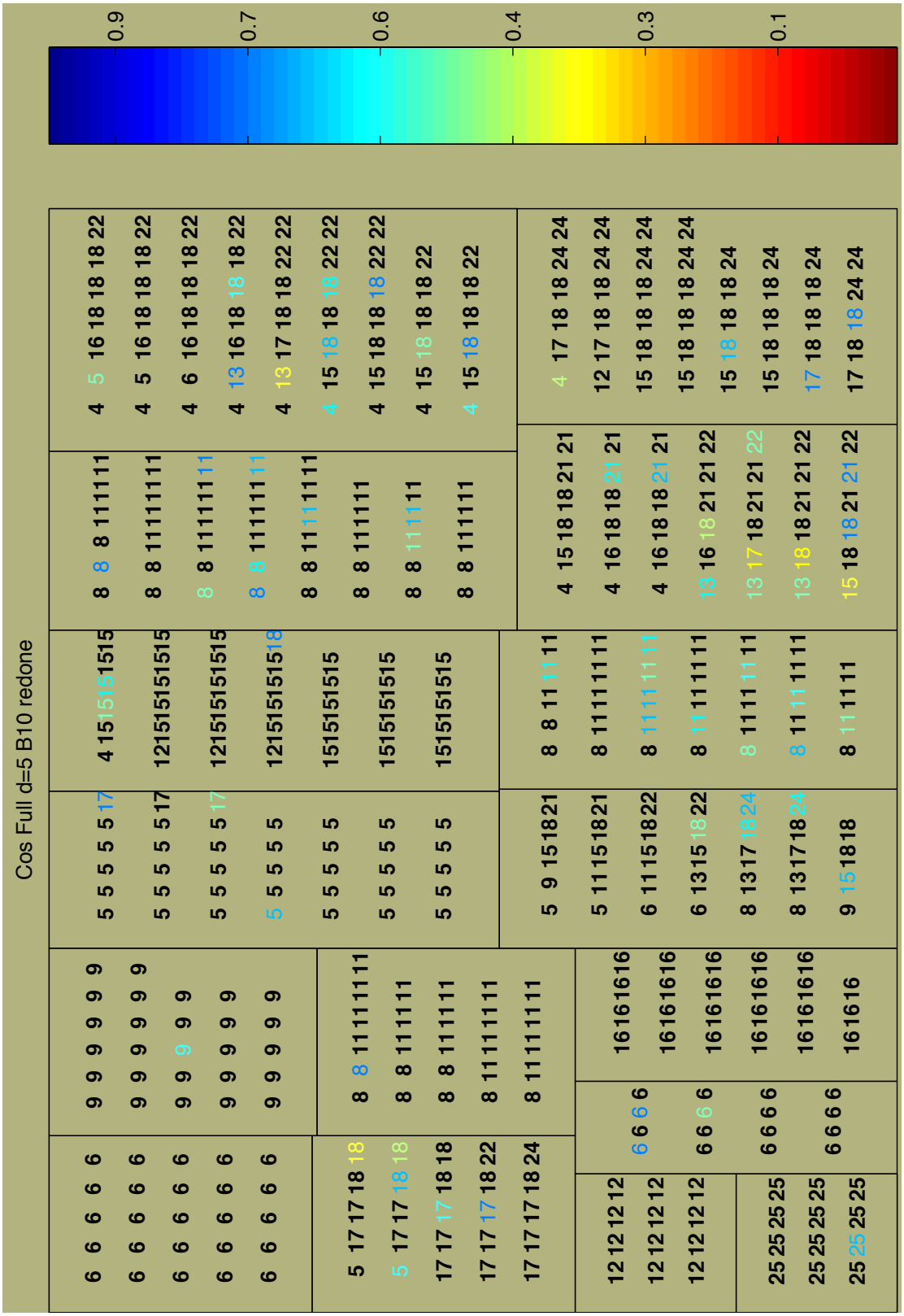

Figure 5. This is the ReClus output from the model-based clustering, where the full text lexicon is used with the Ochiai measure.

References:

Baeza-Yates, Ricardo and Berthier Ribero-Neto, 1999. *Modern Information Retrieval*, ACM Press, New York, NY.

Banfield, J. D. and A. E. Raftery, 1993. 'Model-based Gaussian and non-Gaussian clustering,' *Biometrics*, 49, pp. 803 - 821.

Berry, Michael W., and Murray Browne.1999. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. SIAM, Philadelphia, PA.

Charniak, Eugene. 1996. *Statistical Language Learning*, The MIT Press, Cambridge, MA.

Cover, T. M. and P. E. Hart, 1967. 'Nearest neighbor pattern classification,' *IEEE Transactions on Information Theory*, 13, pp. 21 - 27.

Everitt, B. S. and D. J. Hand, 1981. *Finite Mixture Distributions*, Chapman and Hall, London, UK.

Frakes, W. B. and Ricardo Baeza-Yates, 1992. *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, New Jersey.

Fraley, C. and A. E. Raftery, 1998. 'How many clusters? Which clustering method? - Answers via model-based cluster analysis,' *The Computer Journal*, 41, pp. 578 - 588.

Kimbrell, Roy E., 1988. 'Searching for text? Send an N-Gram!,' *Byte*, May, pp. 297 - 312.

Landauer, Thomas K., Darrell Laham, and Peter Foltz. 1998. Learning human-like knowledge by Singular Value Decomposition: A progress report. In M. I. Jordan, M. J. Kearns and S. A. Solla (Eds.). Advances in Neural Information Processing Systems, 10, pp. 45 - 51. Cambridge. MIT Press.

Martinez, Angel R., 2002. *A Statistical Framework for the Representation of Semantics*, Ph.D. Dissertation, George Mason University (in process).

Nettleton, Dan and T. Bannerjee, 2001. 'Testing the equality of distributions of random vectors with categorical components,' *Computational Statistics and Data Analysis*, 37, pp. 195 - 208.

Porter, M. F., 1980. 'An algorithm for suffix stripping,' *Program*, 14, pp. 130 - 137.

Révézs, Gyorgy. 1983. *Introduction to Formal Languages*, McGraw-Hill Book Company, New York, NY.

Salton, Gerard, Chris Buckley and Maria Smith, 1990. 'On the application of syntactic methodologies,' *Automatic Text Analysis, Information Processing & Management*, 26, pp. 73 - 92.

Tenenbaum, Joshua B., Vin deSilva and John C, Langford, 2000. 'A global geometric framework for nonlinear dimensionality reduction,' *Science*, 290, pp. 2319 - 2323.

Webb, Andrew, 1999. *Statistical Pattern Recognition*, Oxford University Press, Oxford, UK.

Wegman, E. J., 1990. 'Hyperdimensional data analysis using parallel coordinates,' *Journal of the American Statistical Association*, 85, pp. 664 - 675.

Witten, I. H., A. Moffat and T. C. Bell, 1994. *Managing Gigabytes: Compressing ad Indexing Documents and Images*, van Nostrand Reinhold, New York, NY.