Email Surveillance Using Nonnegative Matrix Factorization

Michael W. Berry[*]   Murray Browne[†]

February 25, 2005

Abstract

In this study, we apply a non-negative matrix factorization approach for the extraction and detection of concepts or topics from electronic mail messages. For the publicly released Enron electronic mail collection, we encode sparse term-by-message matrices and use a low rank non-negative matrix factorization algorithm to preserve natural data non-negativity and avoid subtractive basis vector and encoding interactions present in techniques such as principal component analysis. Results in topic detection and message clustering are discussed in the context of published Enron business practices and activities, and benchmarks addressing the computational complexity of our approach are provided. The resulting basis vectors and matrix projections of this approach can be used to identify and monitor underlying semantic features (topics) and message clusters in a general or high-level way without the need to read individual electronic mail messages.

Keywords: electronic mail, Enron collection, non-negative matrix factorization, surveillance, topic detection, constrained least squares.

1 Background

One of the by-products of the Federal Energy Regulatory Commission's (FERC) investigation of Enron was the vast amount of information (electronic mail messages, phone tapes, internal documents) collected towards building a legal case against the global energy corporation. As a matter of public record, this information which initially contained over 1.5 million electronic mail (email) messages was originally posted on FERC's web site [9]. However the original set suffered from document integrity problems and attempts were made to improve the quality of the data and remove some of the sensitive and irrelevant private information. Dr. William Cohen of Carnegie Mellon University took the lead in distributing this improved corpus – known as the Enron

Email Sets. The latest version of the Enron Email Sets[1] (dated – March 2, 2004) contains $517,431$ email messages of 150 Enron employees covering a period from December 1979 through February 2004 with the majority of messages spanning the three years: 1999, 2000, and 2001. It includes messages of some of the top executives of Enron management personnel including founder and Chief Executive Officer (CEO) Ken Lay, president and Chief Operating Officer (COO) Jeff Skilling, and head of trading and later COO, Greg Whalley. Other top executives who played major roles in the day-to-day operations of the corporation are represented as well. They include: Louise Kitchen who developed the Enronline, the corporations in-house trading system, Vince Kaminski head of research, Richard Sanders leader of Enron North America's litigation department and Steve Kean Executive Vice President and Chief of Staff.

In addition to operational logistics of being America's seventh largest company, Enron was faced with many ongoing crises. One involved Enron's development of the Dabhol Power Company (DPC) in the Indian state of Maharashtra, an endeavor awash in years of logistical and political problems. Then there was the deregulation of the California energy market, which led to rolling blackouts during the summer of 2000 - a situation that Enron (and other energy companies) took advantage of financially. By the fall of 2001, Enron's combination of greed, overspeculation, and deceptive accounting practices snowballed into an abrupt collapse. A last minute merger with the Dynegy energy company fell through and Enron filed for Chapter 11 bankruptcy on December 2, 2001 [18]. As expected, The Enron Email Sets reflect this business world ranging from corporate memos to fantasy football picks. The challenge was how to classify this information in a meaningful way.

In Section 2 we discuss one mathematical approach to the extraction of *features* from subcollections of Enron electronic messages – non-negative matrix factoriza-

---

[1] http://www-2.cs.cmu.edu/∼enron

tion. Building upon previous work in topic detection on benchmark collections (with human curated classifications) [21], we apply this *parts*-based factorization approach to topic detection and monitoring of electronic mail messages. Such an application could facilitate the design of future *surveillance* systems in which topics of electronic mail discussions are identified (without literally reading messages) and tracked over time. In Section 3, we describe two particular subsets of the Enron collection that were parsed and analyzed for topic tracking. Section 4 provides illustrations of successful topic detection along with caveats to the use of non-negative factorization in this context. Tracking one particular year (2001) of an Enron subcollection shows which corporate deals and activities dominated the corporation prior to and during its collapse. There is also a discussion of how to distinguish different types of electronic messages. Finally, we conclude our findings and suggest future modeling of the Enron collection in Section 5.

2 Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) has recently been shown to be a very useful technique in approximating high dimensional data where the data are comprised of non-negative components. In a seminal paper published in *Nature* [15], Lee and Seung proposed the idea of using NMF techniques to find a set of basis functions to represent image data where the basis functions enable the identification and classification of intrinsic "parts" that make up the object being imaged by multiple observations. They showed that NMF facilitates the analysis and classification of data from image or sensor articulation databases made up of images showing a composite object in many articulations, poses, or observation views. They also found NMF to be a useful tool in text data mining. In the past few years, several papers have discussed NMF techniques and successful applications to various databases where the data values are non-negative, e.g., [7, 11, 12, 13, 16, 17, 22].

More generally, matrix factorization techniques in data mining fall under the category of vector space methods. Very often databases of interest lead to a very high dimensional matrix representation. Low-rank factorizations not only enable the user to work with reduced dimensional models, they also often facilitate more efficient statistical classification, clustering and organization of data, and lead to faster searches and queries for patterns or trends, e.g., Berry, Drmač, and Jessup [4]. Recently, Xu et al [23] demonstrated that NMF-based indexing outperforms traditional vector space approaches to information retrieval (such as latent semantic indexing) for document clustering on a few benchmark test collections.

NMF is a vector space method used to obtain a representation of data using non-negativity constraints. These constraints can lead to a parts-based representation because they allow only additive, not subtractive, combinations of the original data. This is in contrast to techniques for finding a reduced dimensional representation based on singular value decomposition-type methods such as principal component analysis (PCA) [14]. One major problem with PCA is that the basis vectors have both positive and negative components, and the data are represented as linear combinations of these vectors with positive and negative coefficients. In many applications, however, the negative components contradict physical realities. In particular, term frequencies in text mining are non-negative. In this paper, we survey some popular computational approaches (and their complexities) for NMF in the context of document clustering applications, and demonstrate the use of a *new* hybrid NMF method that can enforce smoothness (or sparsity) constraints on the resulting factor matrices.

2.1 Optimization Problem Given a collection of electronic mail messages expressed as an $m \times n$ term-by-message matrix X, where each column is an m-dimensional non-negative vector of the original collection (n vectors), the standard NMF problem is to find two new reduced-dimensional matrices W and H, in order to approximate the original matrix X by the product WH in terms of some metric. Each column of W contains a *basis vector* while each column of H contains the *weights* needed to approximate the corresponding column in X using the basis from W. The dimensions of matrices W and H are $m \times r$ and $r \times n$, respectively. Usually, the number of columns in the new (basis) matrix W is chosen so that $r \ll n$. Here, the choice of r is generally application dependent, and may also depend upon the characteristics of the particular corpus or database [11].

The usual approach to the NMF problem is to approximate X by computing a pair W and H to minimize the Frobenius norm of the difference $X - WH$. Mathematically, the problem can be stated as follows: Let $X \in R^{m \times n}$ be a data matrix of non-negative entries. Let $W \in R^{m \times r}$ and $H \in R^{r \times n}$ for some positive integer $r < n$. The objective is then to solve the optimization problem

$$(2.1) \qquad \min_{W,H} \|X - WH\|_F^2,$$

subject to $W_{ij} \geq 0$ and $H_{ij} \geq 0$ for each i and j.

Of course the matrices W and H are generally not unique. Conditions resulting in uniqueness in the special case of equality, $X = WH$, have been recently studied by Donoho and Stodden [7], using cone theoretic

46

techniques (See also Chapter 1 in Berman and Plemmons [1]). Algorithms designed to approximate X by solving the minimization problem (2.1) generally begin by initial estimates of the matrices W and H, followed by alternating iterations to improve these estimates.

To explain the non-negative matrix factorization approach used in this study, we briefly review previous methods discussed in the literature.

2.2   Multiplicative Method. A non-negative matrix factorization algorithm of Lee and Seung [15] is based on multiplicative update rules of W and H. We call this scheme the *multiplicative method*, and denote it by **MM**. A formal statement of the method is given below:

MM Algorithm (Lee and Seung)

1. Initialize W and H with non-negative values, and scale the columns of W to unit norm.

2. Iterate for each c, j, and i until convergence or after k iterations:

   (a) $H_{cj} \leftarrow H_{cj} \dfrac{(W^T X)_{cj}}{(W^T W H)_{cj} + \epsilon}$

   (b) $W_{ic} \leftarrow W_{ic} \dfrac{(X H^T)_{ic}}{(W H H^T)_{ic} + \epsilon}$

   (c) Scale the columns of W to unit norm.

Clearly the approximations W and H remain non-negative during the updates. It is generally best to update W and H *simultaneously*, instead of updating each matrix fully before the other. In this case, after updating a row of H, we update the corresponding column of W. In the implementation described in [21], a small positive quantity, say the square root of the machine precision, should be added to the denominators in the approximations of W and H at each iteration step. Setting $\epsilon = 10^{-9}$ will typically suffice.

It is often important to normalize the columns of X in a pre-processing step, and in the algorithm to normalize the columns of the basis matrix W at each iteration. In this case we are optimizing on a unit hypersphere, as the column vectors of W are effectively mapped to the surface of a hypersphere by the repeated normalization.

The computational complexity of Algorithm MM can be shown to be $O(rmn)$ operations per iteration. Additional data (e.g., new electronic mail messages) can either be added directly to the basis matrix W along with a minor modification of H, or else if r is fixed, then further iterations can be applied starting with the current W and H as initial approximations.

Lee and Seung [16] proved that under the MM update rules the distance $\|X - WH\|_F^2$ is monotonically non-increasing. In addition it is invariant if and only if W and H are at a stationary point of the objective function in Eq. (2.1). From the viewpoint of nonlinear optimization, the algorithm can be classified as a diagonally-scaled gradient descent method [11]. Lee and Seung [15] have also provided an additive algorithm. Both the multiplicative and additive algorithms are related to expectation-maximization approaches used in image processing computations such as image restoration, e.g., [20].

2.3   Enforcing Statistical Sparsity. Hoyer [12] has suggested a novel non-negative sparse coding scheme based on ideas from the study of neural networks, and the scheme has been applied to the decomposition of databases into independent feature subspaces by Hyvärinen and Hoyer [13]. Hoyer's method [12] has the important feature of enforcing a statistical sparsity for the weight matrix H, thus enhancing the parts-based representation of the data in W.

Mu, Plemmons and Santago [19] propose a regularization approach that, like Hoyer's method, can be used to enforce statistical sparsity of the weight matrix H. This approach uses a so-called point count regularization scheme in the computations that penalizes the *number* of nonzero entries in H, rather than $\sum_{ij} H_{ij}$, as proposed by Hoyer. Sparsity often leads to a basis representation in W that better represents parts or features of the corpus defined by X [21].

2.4   A Hybrid NMF Approach. We use a hybrid algorithm for NMF that combines some of the better features of available methods. As discussed in [21], the multiplicative algorithm approach can be used to compute an approximation to the basis matrix W at each iterative step. This computation is essentially a matrix version of the gradient descent optimization scheme mentioned earlier. Secondly, we compute the weight matrix H using a constrained least squares (CLS) model as the metric. The purpose is to penalize non-smoothness and non-sparsity in H. This approach bears similarity to those of Hoyer and Mu, Plemmons and Santago. The CLS model is related to the least squares Tikhonov regularization technique commonly used in image restoration [20]. As presented in [21], the algorithm, referred to as **GD-CLS** for *gradient descent with constrained least squares*, is given below:

Algorithm for GD-CLS [21]

1. Initialize W and H with non-negative values, and

scale the columns of W to unit norm.

2. Iterate until convergence or after k iterations:

(a) $W_{ic} \leftarrow W_{ic} \dfrac{(XH^T)_{ic}}{(WHH^T)_{ic} + eps}$, for c and i

(b) Rescale the columns of W to unit norm.

(c) Solve the constrained least squares problem:

$$\min_{H_j}\{\|X_j - WH_j\|_2^2 + \lambda\|H_j\|_2^2\},$$

where the subscript j denotes the jth column, for $j = 1, \ldots, m$. Any negative values in H_j are set to zero. The parameter λ is a regularization value that is used to balance the reduction of the metric

$$\|X_j - WH_j\|_2^2$$

with enforcement of smoothness and sparsity in H.

As done in Algorithm MM, we use a small positive parameter ϵ to avoid dividing by zero or very small numbers and enhance stability in the computations for W in Step 2(a). The numerical approach for solving the constrained least squares problem in Step 2(c) for the columns H_j of H makes use of an algorithm similar to one described in [20] for regularized least squares image restoration.

3 Electronic Mail Subcollections

We have recently tested the effectiveness of the **GD-CLS** algorithm for computing the non-negative matrix factorization of term-by-message matrices derived from the Enron corpus. These matrices were derived from the creation and parsing of two subcollections: INBOX and PRIVATE . Our rationale for creating these two particular subcollections of the raw Enron collection is that INBOX would reflect a standard (perhaps untarnished) repository of all incoming messages to an Enron employee, and PRIVATE would represent personal classifications of messages originally posted to a user's *inbox* folder and then copied or moved to discriminated folders. Although no attempt is made with the **GD-CLS** algorithm to guarantee that messages of the same folder (user-assigned topic) are spanned by similar feature vectors, the semantic interpretation of feature vectors (using the components of the W and H factors) is greatly improved when taking into account the user's original clustering of messages, i.e., message directory path.

3.1 Message Parsing. The INBOX subcollection is comprised all emails contained in the *inbox* folder of all 150 users (or subdirectories) in the raw dataset. Using a 495-term *stoplist* of unimportant terms (or words), the GTP software environment [10] extracted $80,683$ terms from $44,872$ electronic messages. This subcollection reflects 8.7% of the $517,431$ messages in the *raw* Enron collection. With the same stoplist and parsing all users' mail directories with the exception of *all_documents*, *calendar*, *contacts*, *deleted_items discussion_threads*, *inbox*, *notes_inbox*, *sent*, *sent_items*, and *_sent_mail*, GTP extracted $92,133$ terms from $65,033$ messages (29.1% of the raw collection) to define the PRIVATE subcollection. In order to simulate the tracking of topics through an eventful year, say 2001, in the corporate life of Enron, we also created twelve smaller subsets of the PRIVATE subcollection. As depicted in Table 1, all messages sent in a particular month of 2001 were parsed to create twelve separate dictionaries (or sets of terms). As will be discussed in Section 4.3, we use these smaller collections to track topics throughout the year with no accumulation of dictionaries, that is, we apply **GD-CLS** to each corresponding term-by-message matrix separately and extract message clusters (topics) independently. An alternative approach for topic tracking through time would be to update the non-negative matrix factorization with each new month's set of messages. Methods for the efficient updating of Eq. (4.3) are now under consideration (see [21]) and are not in the scope of this work.

In creating the subcollections, all permissible folders are elgible for parsing (no threshold on the number of messages applied) and all message headers are left intact for GTP to process. All terms (or keywords) comprising the resulting dictionary are required to occur at least twice (globally) across the particular subcollection and in two or more messages. In order to define meaningful term-to-message associations for concept discrimination, term weighting is used in the generation of all term-by-message matrices.

3.2 Term Weighting. As explained in [2], a collection of n messages indexed by m terms (or keywords) can be represented as a $m \times n$ term-by-message matrix $X = [x_{ij}]$. Each element or component x_{ij} of the matrix X defines a *weighted* frequency at which term i occurs in message j [3]. We can define

(3.2) $$x_{ij} = l_{ij}g_i d_j,$$

where l_{ij} is the local weight for term i occurring in message j, g_i is the global weight for term i in the subcollection, and d_j is a document normalization factor which specifies whether or not the columns of X (i.e., the documents) are normalized (i.e., have unit length).

Table 1: Counts of messages from the PRIVATE sub-collection that were sent on each month of 2001. The corresponding number of terms parsed for each monthly subset is denoted as well..

Month	Messages	Terms
Jan	3,621	17,888
Feb	2,804	16,958
Mar	3,525	20,305
Apr	4,273	24,010
May	4,261	24,335
Jun	4,324	18,599
Jul	3,077	17,617
Aug	2,828	16,417
Sep	2,330	15,405
Oct	2,821	20,995
Nov	2,204	18,693
Dec	1,489	8,097

Let f_{ij} be the number of times (frequency) that term i appears in message j, and define $p_{ij} = f_{ij}/\sum_j f_{ij}$. Two possible definitions for x_{ij} in Eq. (3.2) are given by Table 2. We use **txx** and **lex** to refer to simple (term) frequency and log-entropy term weighting, respectively.

Table 2: Term weighting schemes used in the parsing of the INBOX and PRIVATE subcollections. No message normalization is applied so that $d_j = 1$ in Eq. (3.2) and base 2 logarithms should be assumed.

Name	Weighting Component	
	Local	Global
txx	Term Frequency	None
	$l_{ij} = f_{ij}$	$g_i = 1$
lex	Logarithmic	Entropy [8]
	$l_{ij} = \log(1 + f_{ij})$	$g_i = 1 +$
		$(\sum_j p_{ij} log(p_{ij}))/\log n$

4 Observations and Results

Figures 1 and 2 illustrate the different cluster sizes obtained from the non-negative matrix factorization of the term-by-message matrix X associated with the PRIVATE collection with log-entropy and simple term frequency weighting, respectively. Here, we approximate

Table 3: **GD-CLS** benchmarks for computing the non-negative factorization in Eq. (4.3), where X is generated from either the INBOX and PRIVATE electronic mail collections. Exactly 50 clusters (topics), which is also the column dimension of the W matrix and row dimension of the H matrix, are generated, and λ is the regularization parameter controlling the sparsity of the matrix H. Time is elapsed CPU time in seconds on a 450MHz (Dual) UltraSPARC-II processor for 100 iterations of **GD-CLS**.

Collection	Mail Messages	Dictionary Terms	λ	Time (sec.)
INBOX	$44,872$	$80,683$	0.1	$1,471$
			0.01	$1,451$
			0.001	$1,521$
PRIVATE	$65,031$	$92,133$	0.1	$51,489$
			0.01	$51,393$
			0.001	$51,562$

the $92,133 \times 65,031$ (sparse) matrix X via

$$(4.3) \qquad X \simeq WH = \sum_{k=1}^{50} W_k H^k ,$$

where W and H are $92,133 \times 50$ and $50 \times 65,031$, respectively, non-negative matrices. W_k denotes the kth column of W, H^k denotes the kth row of the matrix H, and $r = 50$ factors or parts are produced. Clearly, the non-negativity of W and H facilitate a parts-based representation of the matrix X whereby the basis (column) vectors of W or W_k combine to approximate the original columns (messages) of the sparse matrix X. The outer product representation of WH in Eq. (4.3) demonstrates how the rows of H or H^k essentially specify the weights (scalar multiples) of each of the basis vectors needed for each of the 50 parts of the representation. As described in [15], we can interpret the semantic feature represented by a given basis vector W_k by simply sorting (in descending order) its $92,133$ elements and generating a list of the corresponding dominant terms (or keywords) for that feature. In turn, a given row of H having n elements (i.e., H^k) can be used to reveal messages sharing common basis vectors W_k, i.e., similar semantic features or meaning. The columns of H, of course, are the projections of the columns (messages) of X onto the basis spanned by the columns of W. The best choice for the number of parts r (or column rank of W) is certainly problem-dependent or corpus-dependent in this context. However, as discussed in [21] for

49

standard topic detection benchmark collections (with human-curated document clusters) the accuracy of **GD-CLS** for document clustering degrades as the rank r increases or if the sizes of the clusters become greatly imbalanced. Further investigations into the effects of message clustering with larger ranks (beyond 50) are planned.

The association of features (i.e., feature vectors) to the electronic mail messages is accomplished by the nonzeros of each H^k which would be present in the kth *part* of the approximation to X in Eq. (4.3). Each part (or span of W_k) can be used to classify the messages so the sparsity of H greatly affects the diversity of topics with which any particular semantic feature can be associated. In Figures 1 and 2, we show the number of nonzero elements in each H^k of magnitude greater than $row_{max}/10$ for three different choices of λ (namely 0.001, 0.01, and 0.1) in the **GD-CLS** algorithm. Using the rows of the H matrix and a threshold on the nonzero elements to cluster messages, we obtain quite a wide range of cluster sizes. As λ increases, we do not uniformly see a decrease in the cluster sizes as might be expected (due to an expected increase in the sparsity of H). However for most clusters (or rows of H) there is some reduction in the number of elements exceeding the $row_{max}/10$ threshold. The increased sparsity in H for larger values of λ is also reflected in the elapsed CPU times shown in Table 3. For a more thorough assessment of the reduction in statistical sparsity of the matrix H generated by the **GD-CLS** algorithm see [21].

4.1 Topic Extraction. Tables 4 and 5 illustrate some of the extracted topics (i.e., message clusters) as evidenced by large components in the same row of the matrix H (or H^k) generated by **GD-CLS** for the sparse term-by-message matrix associated with the PRIVATE subcollection. The terms corresponding to the 10-largest elements of the particular feature (or part) k are also listed to explain and derive the context of the topic. By feature, we are referring to the k-th column of the matrix factor W or W_k in Eq. (4.3), of course. The seven topics reflected in Table 4 do occur in the parts-based factorization of the matrix X regardless of whether **lex** or **txx** weighting (see Section 3.2) are used by the GTP software environment. The three topics shown in Table 5, however, were extracted only from the use of **txx** weighting. It is interesting to note that the use of a single term-weighting scheme might have a limiting effect on the ability to discern/interpret context of the features produced by a non-negative matrix factorization. Further studies into such effects are needed.

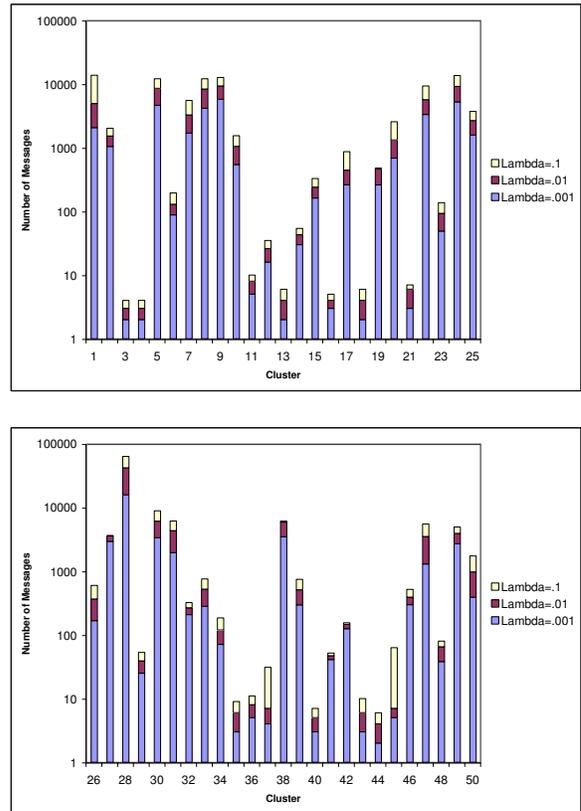In a perfect email surveillance world, each cluster

Figure 1: Size of clusters (number of electronic mail messages) produced by the **GD-CLS** algorithm for the PRIVATE collection. Log-entropy term-weighting is used. Three instances of the regularization parameter ($\lambda = 0.001, 0.01, 0.1$) for controlling the sparsity of the H matrix factor are shown in each graph.

of terms would point to the documents by a specific topic. Although our experiments did not produce such results for every cluster, they did give some indication of what the particular message collection *was about*. With 50 clusters or features produced by **GD-CLS** and deploying both **lex** and **txx** for different instances of a term-by-message matrix X, we analyzed the ten dominant (in magnitude) terms per feature for clues about the content of the collection. The majority of the cluster terms were too vague or too broad to be meaningful, but each variation did reveal clusters that merited further investigation. These clusters had a tendency to have a few proper nouns – words such as *kitchen* (for Louise Kitchen) or company names such as *dynegy* coupled with other more general terms such as *merger* which in the case of *dynegy* and *merger* would point to documents that were referring to the
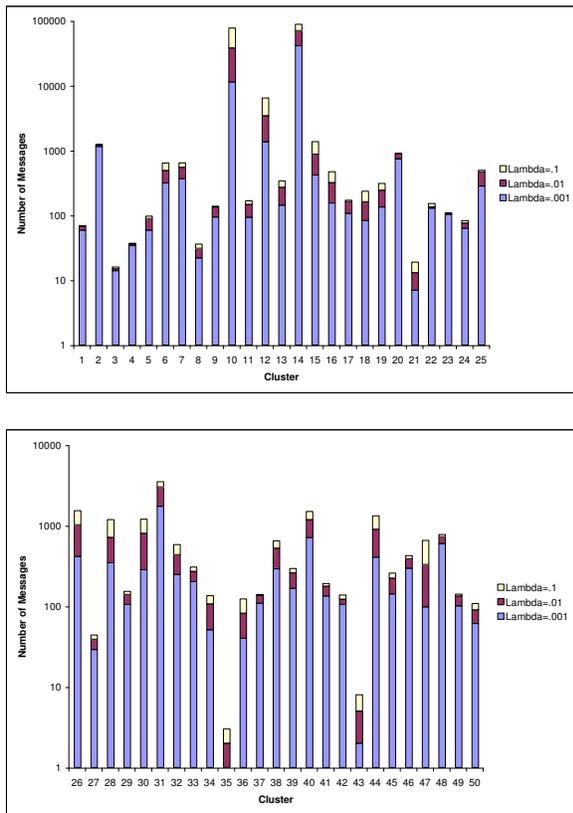
Figure 2: Size of clusters (number of electronic mail messages) produced by the **GD-CLS** algorithm for the PRIVATE collection. Simple term frequency weighting is used. Three instances of the regularization parameter ($\lambda = 0.001, 0.01, 0.1$) for controlling the sparsity of the H matrix factor are shown in each graph.

last minute efforts of Enron to avoid total collapse by merging with the Dynegy corporation. For these type of "meaningful" clusters, we checked to verify that the documents were semantically linked to the terms of the clusters.

The more promising clusters (those that were specific enough to indicate what might be found if one looked at the corresponding documents) were clusters that referred to a median range (say in the hundreds and not thousands) of messages (see Figures 1 and 2). Clusters with only one or several messages were found to be inconclusive. Keep in mind that we are measuring cluster or feature size by the number of row elements in the matrix H with magnitude greater than a specified tolerance (which is $row_{max}/10$ for this study). Conversely, clusters representing thousands of mail messages were unmanageable.

Table 4: Sample clusters (topics) identified by the rows of H or H^k produced by the non-negative matrix factorization (with $\lambda = 0.1$) of the term-by-message matrix X associated with the PRIVATE subcollection and **lex** term-weighting. Exactly $r = 50$ feature vectors (W_k) were generated by the **GD-CLS** algorithm. The ten dominant (having values of largest magnitude) terms for each feature vector are listed for each selected feature (k), and those in **boldface** were judged to be the most descriptive. Cluster size reflects the number of row elements in H^k of magnitude greater than $row_{max}/10$.

Feature Index (k)	Cluster Size	Topic Description	Dominant Terms
10	497	California	ca, **cpuc,** gov, **socalgas,** sempra, org, sce, gmssr, aelaw, ci
23	43	Louise Kitchen named top woman by Fortune	evp, **fortune,** britain, woman, **ceo,** avon, fiorinai, cfo, hewlett, packard
26	231	Fantasy football	game, wr, qb, play, rb, season, injury, updated, fantasy, image
33	233	Texas longhorn football newsletter	UT, orange, longhorn[s], texas, true, truorange, recruiting, oklahoma defensive
34	65	Enron collapse	**partnership[s], fastow,** shares, **sec,** stock, shareholder, investors, equity, **lay**
39	235	Emails about India	**dahhol, dpc, india, mseb, maharashtra,** indian, lenders, delhi, foreign, minister
46	127	Enron collapse	dow, debt, reserved, wall, copyright jones, cents, analysts, reuters, spokesman

51

For example, a cluster with the terms *power, california, electricity, demand* represented $2,253$ documents (a similar California cluster had $8,500$ messages) which is so general that it is of limited use. With this is mind, we made another pass looking at each of the clusters that represented anywhere from 10 to 500 messages even if their terms were initially seemed vague.

One example of a meaningful cluster that seemed too vague at first was the cluster associated with feature index $k = 23$ (see Table 4). This feature spanned such terms as: *evp, fortune, britain, women, ceo, avon, fiorina, cfo, hewlett,* and *packard*. But the cluster defined by the dominant components of H^{23} was composed of 43 messages and thus merited further investigation. A look at those documents revealed a set of electronic mail messages that referred to Louise Kitchen's selection in *Fortune*'s 2001 List of the Fifty Most Powerful Women in Business. The messages even included congratulatory notes from her Enron colleagues.

Perhaps one of the most revealing clusters of this series of experiments, were the football-related clusters. Not only did the clusters reveal which messages (and their participants) were linked to football, but it was able to differentiate between fantasy football leagues, which are typically associated with professional teams, and the University of Texas Longhorn football team.

In the three topics that were unique to the **txx** weighting, the cluster of messages associated with feature $k = 16$ in Table 5 is merely a list of rampant database error messages that were forwarded to a user. The first cluster in that table ($k = 2$) refers to a series of memos about preparing for a possible investigation from a California state senator and the third cluster ($k = 40$) focuses on various gas and oil contracts.

4.2 Message Size. One problem in working with such a volume of emails is that the clusters can be influenced by news wire feeds and other automatically generated content. When examining the messages of each cluster, a message corresponding to the largest component of H^k was usually a news wire feed. For example, with feature $k = 39$ in Table 4 we find that 4.50 is the highest value associated with any message in the cluster. As expected, checking the message reveals a $1,700$-line *Wall Street Journal* news wire article on Dabhol. Component values of H^k for a specific cluster k can also help reveal which messages are news feeds (of little surveillance value) and which messages may be smaller emails with more concise content. For example, in feature $k = 39$, one message identified by a component value (in the k-th row of H or H^k) of just 0.9 is a short message from Vince Kaminski to Jeff

Skilling but it belongs in the India topic cluster because the message strategizes about India. The ability to distinguish between large messages such as news feeds and smaller more personal messages can be gauged by the type of term-weighing scheme (e.g., **lex** or **txx**) deployed. See [8] and [2] for more details on specific attempts to take document (or message) length into account for term-weighting.

Ironically, in the early stages of our results assessment it was the prevalence (and frustration) of the large news wire stories in the Enron INBOX subcollection that prompted us to concentrate more on the PRIVATE subcollection. Also, as mentioned earlier, the PRIVATE subcollection of emails from the Enron Email Sets represents a larger portion of the collection of over 65,000 messages as compared to only approximately $45,000$ messages for the INBOX subcollection. One could also make the argument that in general the messages comprising PRIVATE subcollection were more important to a Enron employee because he or she had to at least evaluate the content of the messages before categorizing them (i.e., moving them to folders).

Table 5: Selected clusters (topics) identified by the rows of H or H^k produced by the non-negative matrix factorization of the term-by-message matrix X associated with the PRIVATE subcollection and **txx** term-weighting. Exactly $r = 50$ feature vectors (W_k) were generated by the **GD-CLS** algorithm (with $\lambda = 0.1$). The ten dominant (having values of largest magnitude) terms for each feature vector are listed for each selected feature (k), and those in **boldface** were judged to be the most descriptive. Cluster size reflects the number of row elements in H^k of magnitude greater than $row_{max}/10$.

Feature Index (k)	Cluster Size	Topic Description	Dominant Terms
2	13	Dunn investigation; document retention policy	documents, committee, subpoena, intended, **brobeck**, senate, records, recipient, email, section
16	156	Database error messages	database, dbcaps97data, **davis**, unknown, alias, pete, date, bill, mark, error
40	311	Gas contracts	gas, natural, oil, pipeline, contract, storage, **el**, prices, **paso**, daily

4.3 Temporal Monitoring. Because the calendar year 2001 comprised the largest volume of electronic mail of any single year of the Enron subcollections considered, we examined the performance of a rank $r = 50$ non-negative factorization (with **lex** term-weighting and $\lambda = 0.1$) on twelve specific subsets of the PRIVATE subcollection. Namely, we isolated those electronic messages sent in each month of the calendar year 2001. We looked at how previously defined topics such as California, India, the bankruptcy after the Dynegy merger fell through, and both football topics (fantasy and college) were represented on a month to month basis. Figure 3 illustrates how clusters/topics identified by the non-negative matrix factorization can be traced through time. The results were consistent with what might expect given the history of the Enron Corporation in 2001. The year began with California Governor Gray Davis calling for an investigation of Enron in light of the 2000 California Energy crisis and it was an ongoing topic throughout the year. To a lesser degree, the discussion and legal battles involving the Dabhol Power Company were also consistently present throughout the year. Perhaps a more poignant example of how the **GD-CLS**-generated clusters reflect timeliness is with the topic of the Dynegy merger and subsequent bankruptcy of Enron. These clusters came to the forefront in the fourth quarter of the year which coincided with Enron's final attempts in November to save itself by merging with Dynegy. The football clusters also demonstrate the ability of the clusters to reflect chronological events. As one would expect, college football dominated in the fall and fantasy (professional) football came on strong in December. The most noteworthy aspect of the temporal monitoring is that the process even identified a cluster of Texas football messages present in May of 2001 (perhaps reflecting the university's spring football practices).

Although, the **GD-CLS**-derived models were unable to generate clusters of very specific topics (something that would be of great value for email surveillance), the resulting parts-based factorizations do give a sense of "aboutness" to the Enron world of international energy management and some general direction on where specific documents on certain topics may be found.

5 Concluding Remarks

We have demonstrated how the **GD-CLS** algorithm for computing the non-negative matrix factorization can be used for extracting and tracking topics of discussion from corporate email. This algorithm effectively computes a parts-based approximation $X \simeq WH$ of a sparse term-by-message matrix X in which the quality
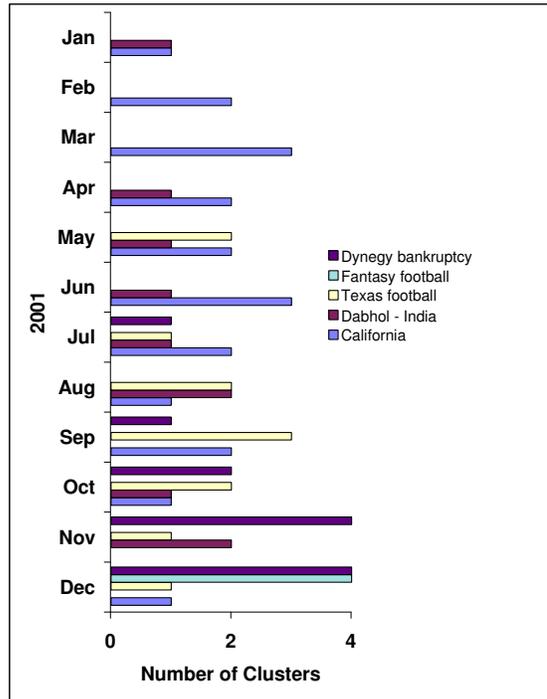

Figure 3: Number of instances of detectable topics for the calendar year 2001 using $r = 50$ features produced by the **GD-CLS** algorithm. Twelve subsets of the PRIVATE collection (one per month) were parsed with each subset comprising all electronic messages sent during the corresponding month of 2001. Log-entropy (or **lex**) term-weighting and the regularization parameter $\lambda = 0.1$ was used for each run of the **GD-CLS** algorithm.

of approximation (error reduction) can be enhanced by an enforcement of smoothness and sparsity in the non-negative matrix H. Although little or no information was extracted to potentially expose fraudulent actions or behaviors of Enron employees, we have demonstrated how a parts-based representation of corporate electronic mail (e.g., Enron) can facilitate the *observation* of electronic message discussions without requiring human intervention or the reading of individual messages. Such surveillance enables corporate leaders (say managers or supervisors) to monitor discussions without the need to isolate or perhaps incriminate individual employees. In this way, factors such as company morale, employees' feedback to policy decisions, and extracurricular activities may eventually be tracked.

With respective to the **GD-CLS** algorithm, further work is needed in exploring the effects of different term

weighting schemes (for X) on the quality of the basis vectors W_k. How document (or message) clustering changes with different column ranks in the matrix W should be considered as well.

References

[1] A. Berman and R. Plemmons. *Non-Negative Matrices in the Mathematical Sciences*, SIAM Press Classics Series, Philadelphia, 1994.

[2] M. Berry and M. Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, SIAM, Philadelphia, 1999.

[3] M. Berry, S. Dumais, and G. O'Brien "Using Linear Algebra for Intelligent Information Retrieval", *SIAM Review*, Vol. 37, No. 4, pp. 573-595, 1995.

[4] M. Berry, Z. Drmač, and E. Jessup. "Matrices, Vector Spaces, and Information Retrieval", *SIAM Review*, Vol. 41, No. 2, pp. 335-362, 1999.

[5] *Concise Columbia Encyclopedia*. Columbia University Press, New York, Second Edition, 1989.

[6] M. Cooper and J. Foote, "Summarizing Video using Non-Negative Similarity Matrix Factorization", *Proc. IEEE Workshop on Multimedia Signal Processing* St. Thomas, US Virgin Islands, 2002.

[7] D. Donoho and V. Stodden. "When does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?", preprint, Department of Statistics, Stanford University, 2003.

[8] S. Dumais, "Improving the Retrieval of Information from External Sources", *Behavior Research Methods, Instruments, & Computers*, Vol. 23, No. 2, pp. 229-236, 1991.

[9] T. Grieve, "The Decline and Fall of the Enron Empire", *Slate*, October 14, 2003, `http://www.salon.com/news/feature/2003/10/14/enron/index_np.html`.

[10] J.T. Giles, L. Wo, and M.W. Berry. "GTP (General Text Parser) Software for Text Mining", in *Statistical Data Mining and Knowledge Discovery*, H. Bozdogan (Ed.), CRC Press, Boca Raton, (2003), pp. 455-471.

[11] D. Guillamet and J. Vitria. "Determining a Suitable Metric when Using Non-Negative Matrix Factorization", *16th International Conference on Pattern Recognition (ICPR'02)*, Vol. 2, Quebec City, QC, Canada, 2002.

[12] P. Hoyer. "Non-Negative Sparse Coding", *Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing)*, Martigny, Switzerland, 2002.

[13] A. Hyvärinen and P. Hoyer. "Emergence of Phase and Shift Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces", *Neural Computation*, Vol. 12, pp. 1705-1720, 2000.

[14] I. Jolliffe. *Principle Component Analysis*, 2nd Ed., Springer Series in Statistics, Springer-Verlag, New York, 2002.

[15] D. Lee and H. Seung. "Learning the Parts of Objects by Non-Negative Matrix Factorization", *Nature*, Vol. 401, pp. 788-791, 1999.

[16] D. Lee and H. Seung. "Algorithms for Non-Negative Matrix Factorization", *Advances in Neural Processing*, 2000.

[17] W. Liu and J. Yi. "Existing and New Algorithms for Non-negative Matrix Factorization", preprint, Computer Sciences Department, University of Texas at Austin, 2003.

[18] B. Mclean and P. Elkind. *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron*, Portfolio, 2003.

[19] Z. Mu, R. Plemmons and P. Santago. "Iterative Ultrasonic Signal and Image Deconvolution for Estimating the Complex Medium Response", preprint, submitted to *IEEE Transactions on Ultrasonics and Frequency Control*, 2003.

[20] S. Prasad, T. Torgersen, V. Pauca, R. Plemmons, and J. van der Gracht. "Restoring Images with Space Variant Blur via Pupil Phase Engineering", Optics in Info. Systems, Special Issue on Comp. Imaging, SPIE Int. Tech. Group Newsletter, Vol. 14, No. 2, pp. 4-5, 2003.

[21] F. Shanaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons. "Document Clustering Using Nonnegative Matrix Factorization", *Information Processing & Management*, 2005, to appear.

[22] S. Wild, J. Curry and A. Dougherty. "Motivating Non-Negative Matrix Factorizations", *Proceedings of the Eighth SIAM Conference on Applied Linear Algebra*, Williamsburg, VA, July 15-19, 2003. See `http://www.siam.org/meetings/la03/proceedings/`.

[23] W. Xu, X. Liu, and Y. Gong. "Document-Clustering based on Non-Negative Matrix Factorization", *Proceedings of SIGIR'03*, July 28 - August 1, Toronto, CA, pp. 267-273, 2003.