Text Mining Approaches for Email Surveillance

Document Space Workshop, IPAM/UCLA

Michael W. Berry and Murray Browne

Department of Computer Science, UTK

January 23, 2006

Enron Background

Non-Negative Matrix Factorization (NMF)

Electronic Mail Surveillance

Conclusions and References

Email Collection

- By-product of the FERC investigation of Enron (originally contained 15 million email messages).

Email Collection

- By-product of the FERC investigation of Enron (originally contained 15 million email messages).
- This study used the improved corpus known as the Enron Email set, which was edited by Dr. William Cohen at CMU.

Email Collection

- By-product of the FERC investigation of Enron (originally contained 15 million email messages).
- This study used the improved corpus known as the Enron Email set, which was edited by Dr. William Cohen at CMU.
- This set had over 500,000 email messages. The majority were sent in the 1999 to 2001 timeframe.

Enron Background

- Email Collection
- Historical Events

Non-Negative Matrix Factorization (NMF)

Electronic Mail Surveillance

Conclusions and References

Enron Historical 1999-2001

- Ongoing, problematic, development of the Dabhol Power Company (DPC) in the Indian state of Maharashtra.

Enron Historical 1999-2001

- Ongoing, problematic, development of the Dabhol Power Company (DPC) in the Indian state of Maharashtra.
- Deregulation of the Calif. energy industry, which led to rolling electricity blackouts in the summer of 2000 (and subsequent investigations).

Enron Historical 1999-2001

- Ongoing, problematic, development of the Dabhol Power Company (DPC) in the Indian state of Maharashtra.
- Deregulation of the Calif. energy industry, which led to rolling electricity blackouts in the summer of 2000 (and subsequent investigations).
- Revelation of Enron's deceptive business and accounting practices that led to an abrupt collapse of the energy colossus in October, 2001; Enron filed for bankruptcy in December, 2001.

NMF Origins

- NMF (Nonnegative Matrix Factorization) can be used to approximate high-dimensional data having nonnegative components.

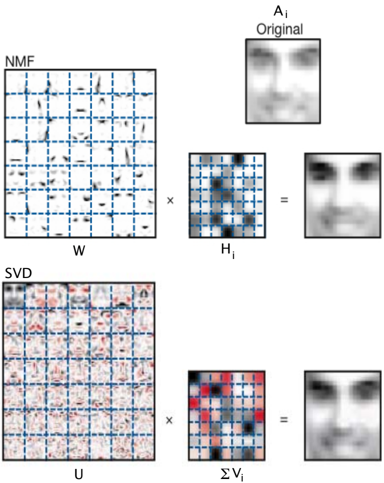
NMF Origins

- NMF (Nonnegative Matrix Factorization) can be used to approximate high-dimensional data having nonnegative components.
- Lee and Seung (1999) demonstrated its use as a *sum-by-parts* representation of image data in order to both identify and classify image *features*.

NMF Origins

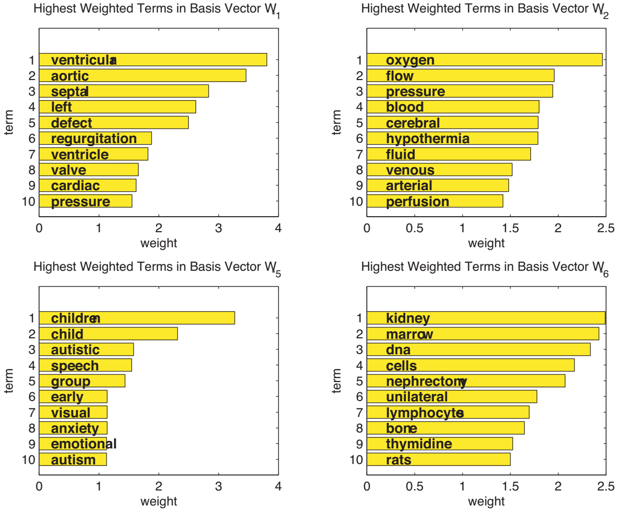
- NMF (Nonnegative Matrix Factorization) can be used to approximate high-dimensional data having nonnegative components.
- Lee and Seung (1999) demonstrated its use as a *sum-by-parts* representation of image data in order to both identify and classify image *features*.
- [Xu et al., 2003] demonstrated how NMF-based indexing could outperform SVD-based Latent Semantic Indexing (LSI) for some information retrieval tasks.

NMF for Image Processing



Sparse NMF verses Dense SVD Bases; Lee and Seung (1999)

NMF for Text Mining (Medlars)



Interpretable NMF feature vectors; Langville et al. (2006)

Derivation

- Given an $m \times n$ term-by-message (sparse) matrix X.

Derivation

- Given an $m \times n$ term-by-message (sparse) matrix X.
- Compute two reduced-dim. matrices W,H so that $X \simeq WH$; W is $m \times r$ and H is $r \times n$, with $r \ll n$.

Derivation

- Given an $m \times n$ term-by-message (sparse) matrix X.
- Compute two reduced-dim. matrices W,H so that $X \simeq WH$; W is $m \times r$ and H is $r \times n$, with $r \ll n$.
- **Optimization problem**:

$$\min_{W,H} \|X - WH\|_F^2,$$

subject to $W_{ij} \geq 0$ and $H_{ij} \geq 0, \ \forall i, j$.

Derivation

- Given an $m \times n$ term-by-message (sparse) matrix X.
- Compute two reduced-dim. matrices W,H so that $X \simeq WH$; W is $m \times r$ and H is $r \times n$, with $r \ll n$.
- **Optimization problem**:

$$\min_{W,H} \|X - WH\|_F^2,$$

subject to $W_{ij} \geq 0$ and $H_{ij} \geq 0$, $\forall i, j$.

- **General approach**: construct initial estimates for W and H and then improve them via alternating iterations.

Multiplicative Method (MM)

- Multiplicative update rules for W and H (Lee and Seung, 1999):
  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:
     - 2.1 $H_{cj} \leftarrow H_{cj} \dfrac{(W^T X)_{cj}}{(W^T WH)_{cj} + \epsilon}$
     - 2.2 $W_{ic} \leftarrow W_{ic} \dfrac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$
     - 2.3 Scale the columns of W to unit norm.

Multiplicative Method (MM)

- Multiplicative update rules for W and H (Lee and Seung, 1999):
  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:
     2.1 $H_{cj} \leftarrow H_{cj} \dfrac{(W^T X)_{cj}}{(W^T WH)_{cj} + \epsilon}$
     2.2 $W_{ic} \leftarrow W_{ic} \dfrac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$
     2.3 Scale the columns of W to unit norm.
- Setting $\epsilon = 10^{-9}$ will suffice [Shahnaz et al., 2006].

Normalization, Complexity, and Convergence

- Important to normalize X initially and the basis matrix W at each iteration.

Normalization, Complexity, and Convergence

- Important to normalize X initially and the basis matrix W at each iteration.
- When optimizing on a unit hypersphere, the column (or feature) vectors of W, denoted by W_k, are effectively mapped to the surface of the hypersphere by repeated normalization.

Normalization, Complexity, and Convergence

- Important to normalize X initially and the basis matrix W at each iteration.
- When optimizing on a unit hypersphere, the column (or feature) vectors of W, denoted by W_k, are effectively mapped to the surface of the hypersphere by repeated normalization.
- MM implementation of NMF requires $\mathcal{O}(rmn)$ operations per iteration; Lee and Seung (1999) proved that $\|X - WH\|_F^2$ is monotonically non-increasing with MM.

Normalization, Complexity, and Convergence

- Important to normalize X initially and the basis matrix W at each iteration.
- When optimizing on a unit hypersphere, the column (or feature) vectors of W, denoted by W_k, are effectively mapped to the surface of the hypersphere by repeated normalization.
- MM implementation of NMF requires $\mathcal{O}(rmn)$ operations per iteration; Lee and Seung (1999) proved that $\|X - WH\|_F^2$ is monotonically non-increasing with MM.
- From a nonlinear optimization perspective, MM/NMF can be considered a **diagonally-scaled gradient descent method**.

Hoyer's Method

- From neural network applications, Hoyer (2002) enforced statistical sparsity for the weight matrix H in order to enhance the parts-based data representations in the matrix W.

Hoyer's Method

- From neural network applications, Hoyer (2002) enforced statistical sparsity for the weight matrix H in order to enhance the parts-based data representations in the matrix W.
- Mu et al. (2003) suggested a regularization approach to achieve statistical sparsity in the matrix H: **point count regularization**; penalize the *number* of nonzeros in H rather than $\sum_{ij} H_{ij}$.

Hoyer's Method

- From neural network applications, Hoyer (2002) enforced statistical sparsity for the weight matrix H in order to enhance the parts-based data representations in the matrix W.
- Mu et al. (2003) suggested a regularization approach to achieve statistical sparsity in the matrix H: **point count regularization**; penalize the *number* of nonzeros in H rather than $\sum_{ij} H_{ij}$.
- Goal of increased sparsity – better representation of *parts* or *features* spanned by the corpus (X) [Shahnaz et al., 2006].

GD-CLS – Hybrid Approach

- First use MM to compute an approximation to W for each iteration – a gradient descent (**GD**) optimization step.

GD-CLS – Hybrid Approach

- First use MM to compute an approximation to W for each iteration – a gradient descent (**GD**) optimization step.
- Then, compute the weight matrix H using a constrained least squares (**CLS**) model to penalize non-smoothness (i.e., non-sparsity) in H – common Tikohonov regularization technique used in image processing (Prasad et al., 2003).

GD-CLS – Hybrid Approach

- First use MM to compute an approximation to W for each iteration – a gradient descent (**GD**) optimization step.
- Then, compute the weight matrix H using a constrained least squares (**CLS**) model to penalize non-smoothness (i.e., non-sparsity) in H – common Tikohonov regularization technique used in image processing (Prasad et al., 2003).
- Convergence to a non-stationary point evidenced (but no formal proof given to date).

GD-CLS Algorithm

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

GD-CLS Algorithm

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:
   2.1 $W_{ic} \leftarrow W_{ic} \dfrac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$, for c and i
   2.2 Rescale the columns of W to unit norm.
   2.3 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 j^{th} column, for $j = 1, \ldots, m$.

GD-CLS Algorithm

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:
   2.1 $W_{ic} \leftarrow W_{ic} \dfrac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$, for c and i
   2.2 Rescale the columns of W to unit norm.
   2.3 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 j^{th} 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 [Shahnaz et al., 2006].

INBOX Collection

- Parsed *inbox* folder of all 150 accounts (users) via **GTP** (General Text Parser); 495-term stoplist used and extracted terms must appear in more than 1 email and more than once globally.

PRIVATE Collection

- Parsed all mail directories (of all 150 accounts) with the exception of `all_documents`, `calendar`, `contacts`, `deleted_items`, `discussion_threads`, `inbox`, `notes_inbox`, `sent`, `sent_items`, and `_sent_mail`; 495-term stoplist used and extracted terms must appear in more than 1 email and more than once globally.

PRIVATE Collection

- Parsed all mail directories (of all 150 accounts) with the exception of `all_documents`, `calendar`, `contacts`, `deleted_items`, `discussion_threads`, `inbox`, `notes_inbox`, `sent`, `sent_items`, and `_sent_mail`; 495-term stoplist used and extracted terms must appear in more than 1 email and more than once globally.
- Distribution of messages sent in the year 2001:

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

Term Weighting Schemes

- For $m \times n$ term-by-message matrix $X = [x_{ij}]$, define

$$x_{ij} = l_{ij} g_i d_j,$$

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 (set $d_j = 1$).

Term Weighting Schemes

- For $m \times n$ term-by-message matrix $X = [x_{ij}]$, define

$$x_{ij} = l_{ij} g_i d_j,$$

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 (set $d_j = 1$).

- Schemes used in parsing INBOX and PRIVATE subcollections:

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

Computational Complexity

- Rank-50 NMF ($X \simeq WH$) computed on a 450MHz (Dual) UltraSPARC-II processor using 100 iterations:

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

PRIVATE with Log-Entropy Weighting

- Identify rows of H from $X \simeq WH$ or H^k with $\lambda = 0.1$; $r = 50$ feature vectors (W_k) generated by GD-CLS:

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, fiorina, cfo, hewlett, packard
26	231	Fantasy football	game, wr, qb, play, rb, season, injury, updated, fantasy, image

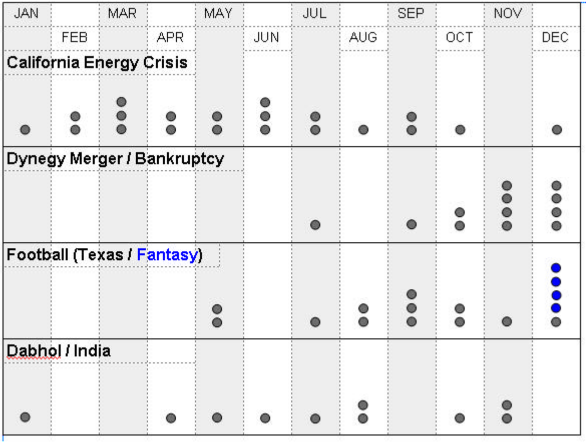
(Cluster size $\equiv$ no. of H^k elements $> row_{max}/10$)

PRIVATE with Log-Entropy Weighting

- Additional topic clusters of significant size:

Feature Index (k)	Cluster Size	Topic Description	Dominant Terms
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	**dabhol**, **dpc**, **india**, **mseb**, **maharashtra**, indian, lenders, delhi, foreign, minister

2001 Topics Tracked by GD-CLS



$r = 50$ features, **lex** term weighting, $\lambda = 0.1$

Two Penalty Term Formulation

- Introduce smoothing on W_k (feature vectors) in addition to H^k:

$$\min_{W,H}\{\|X - WH\|_F^2 + \alpha\|W\|_F^2 + \beta\|H\|_F^2\},$$

where $\|\cdot\|_F$ is the Frobenius norm.

Two Penalty Term Formulation

- Introduce smoothing on W_k (feature vectors) in addition to H^k:

$$\min_{W,H}\{\|X - WH\|_F^2 + \alpha\|W\|_F^2 + \beta\|H\|_F^2\},$$

where $\|\cdot\|_F$ is the Frobenius norm.

- Constrained NMF (CNMF) iteration [Piper et al., 2004]:

$$H_{cj} \leftarrow H_{cj}\frac{(W^TX)_{cj} - \beta H_{cj}}{(W^TWH)_{cj} + \epsilon}$$

$$W_{ic} \leftarrow W_{ic}\frac{(XH^T)_{ic} - \alpha W_{ic}}{(WHH^T)_{ic} + \epsilon}$$

Term Distribution in Feature Vectors

Terms	Wt	Lambda			Alpha			Topics
		0.1	0.01	0.001	0.1	0.01	0.001	
Blackouts	0.508				4	6	4	Cal
Stocks	0.511				2			Collapse
UT	0.517				2			Texasfoot
Chronicle	0.523				3	2	3	
Indian	0.527				2			India
Fastow	0.531				5	3	4	Collapse
Gas	0.531					2	2	
CFO	0.556				2		2	Kitchen
Californians	0.557					3		Cal
Solar	0.570				2			
Partnerships	0.576				6	2	5	Collapse
Workers	0.577					3	2	
Maharashtra	0.591				2		2	India
Mseb	0.605				2			India
Beach	0.611			2				
Ljm	0.621				3		3	Collapse
Tues	0.626		2	2				
IPPS	0.644			2		2		Cal
Rebates	0.647					2		
Ljm2	0.688				2		2	Collapse

British National Corpus (BNC) Noun Conservation

- In collaboration with P. Keila and D. Skillicorn (Queens Univ.)
- 289,695 email subset (all mail folders - not just private)
- Smoothing solely applied to NMF W matrix ($\alpha = 0.001, 0.01, 0.1, 0.25, 0.50, 0.75, 1.00$ with $\beta = 0$)
- Log-entropy term weighting applied to the term-by-message matrix X
- Monitor top ten nouns for each feature vector (ranked by descending component values) and extract those appearing in two or more features; topics assigned manually.

BNC Noun Distribution in Feature Vectors

Noun	GF	Entropy	Alpha							Topic
			0.001	0.01	0.1	0.25	0.50	0.75	1.00	
Waxman	680	0.424	2		2	2	2	2		Downfall
Lieberman	915	0.426	2	2	2	2			2	Downfall
Scandal	679	0.428	2				2		2	Downfall
Nominee(s)	544	0.436		4	3	2		2	2	
Barone	470	0.437	2	2	2				2	Downfall
MEADE	456	0.437							2	Downfall
Fichera	558	0.438	2			2				California blackout
Prabhu	824	0.445	2	2	2	2		2	2	India-strong
Tata	778	0.448							2	India-weak
Rupee(s)	323	0.452	3	4	4	4	3	4	2	India-strong
Soybean(s)	499	0.455	2	2	2	2	2	2	2	
Rushing	891	0.486	2	2	2					Football - college
Dlrs	596	0.487							2	
Janus	580	0.488	2	3				2	3	India-weak
BSES	451	0.498	2	2					2	India-weak
Caracas	698	0.498						2		
Escondido	326	0.504	2			2				California/Blackout
Promoters	180	0.509	2							Energy/Scottish
Aramco	188	0.550	2							India-weak
DOORMAN	231	0.598		2						Bawdy/Real Estate

Enron Background

Non-Negative Matrix Factorization (NMF)

Electronic Mail Surveillance

Conclusions and References
- Conclusions
- Future Work
- References

Conclusions

- GD-CLS Algorithm can effectively produce a *parts-based* approximation $X \simeq WH$ of a sparse term-by-message matrix X.

Conclusions

- GD-CLS Algorithm can effectively produce a *parts-based* approximation $X \simeq WH$ of a sparse term-by-message matrix X.
- Smoothing on the features matrix (W) as opposed to the weight matrix H forces more reuse of higher weighted terms.

Conclusions

- GD-CLS Algorithm can effectively produce a *parts-based* approximation $X \simeq WH$ of a sparse term-by-message matrix X.
- Smoothing on the features matrix (W) as opposed to the weight matrix H forces more reuse of higher weighted terms.
- Surveillance systems based on GD-CLS could be used to monitor discussions without the need to isolate or perhaps incriminate individual employees.

Conclusions

- GD-CLS Algorithm can effectively produce a *parts-based* approximation $X \simeq WH$ of a sparse term-by-message matrix X.
- Smoothing on the features matrix (W) as opposed to the weight matrix H forces more reuse of higher weighted terms.
- Surveillance systems based on GD-CLS could be used to monitor discussions without the need to isolate or perhaps incriminate individual employees.
- Potential applications include the monitoring/tracking of company morale, employee feedback to policy decisions, and extracurricular activities

Future Work

- Further work needed in determining effects of alternative term weighting schemes (for X) and choices of control parameters (α, β, λ) on quality of the basis vectors W_k.

Future Work

- Further work needed in determining effects of alternative term weighting schemes (for X) and choices of control parameters (α, β, λ) on quality of the basis vectors W_k.
- How does document (or message) clustering change with different column ranks (r) in the matrix W?

Future Work

- Further work needed in determining effects of alternative term weighting schemes (for X) and choices of control parameters (α, β, λ) on quality of the basis vectors W_k.
- How does document (or message) clustering change with different column ranks (r) in the matrix W?
- Use MailMiner and similar text mining software to produce a topic **annotated** Enron email subset for the public domain.

Future Work

- Further work needed in determining effects of alternative term weighting schemes (for X) and choices of control parameters (α, β, λ) on quality of the basis vectors W_k.
- How does document (or message) clustering change with different column ranks (r) in the matrix W?
- Use MailMiner and similar text mining software to produce a topic **annotated** Enron email subset for the public domain.
- Explore use of NMF for automated gene classification; **Semantic Gene Organizer** (K. Heinrich, PhD Thesis 2006)

For Further Reading

- F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons.
  Document Clustering Using Nonnegative Matrix Factorization.
  Information Processing & Management 42(2), 2006, pp. 373-386.
- J. Piper, P. Pauca, R. Plemmons, and M. Giffin.
  Object Characterization from Spectral Data using Independent Component Analysis and Information Theory.
  Proc. AMOS Technical Conference, Maui, HI, September 2004.
- 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.

SMD06 Text Mining Workshop

TEXT MINING 2006

Hyatt Regency Bethesda
Bethesda, Maryland
April 22, 2006

to be held in conjunction with

Sixth SIAM International Conference on Data Mining (SDM 2006)

and also in conjunction with SIAM's Link Analysis, Counterterrorism Security Workshop, which is also being held on April 22, 2006.

Website: http://www.cs.utk.edu/tmw06