Using Nonnegative Matrix and Tensor Factorizations for Topic and Scenario Detection and Tracking

Math and Stat Colloquim, Utah State University

Michael W. Berry

Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville

February 19, 2009

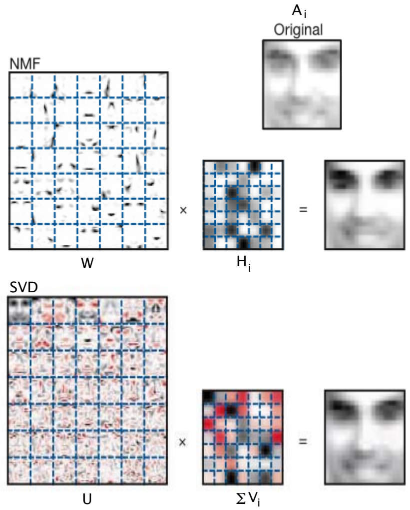
Collaborators

- Brett Bader (Sandia National Labs)
- Murray Browne (Turner Broadcasting System, Inc.)
- Pau'l Pauca, Bob Plemmons, Brian Lamb (Wake Forest)
- Amy Langville (College of Charleston)
- David Skillicorn, Parambir Keila (Queens U.)
- Stratis Gallopoulos, Ioannis Antonellis (U. Patras)

NNMF Origins

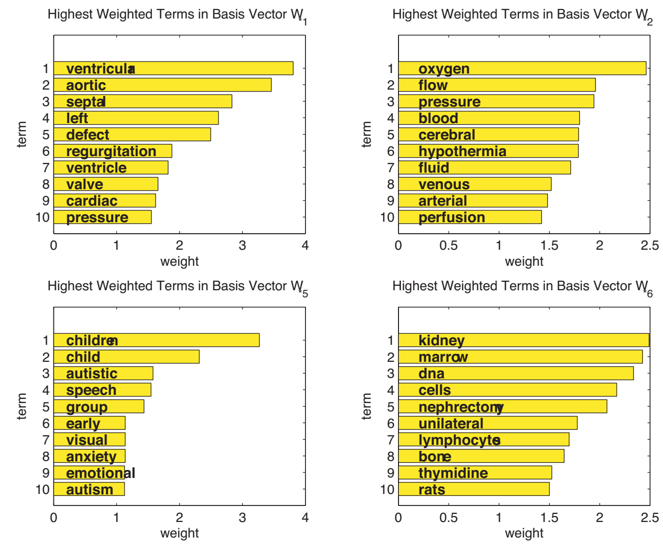
- NNMF (Nonnegative Matrix Factorization) can be used to approximate high-dimensional data having nonnegative components.
- Lee and Seung (1999) demonstrated its use as a *sum-by-parts* representation of image data in order to both identify and classify image *features*.
- Xu et al. (2003) demonstrated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

NNMF for Image Processing



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

NNMF for Text Mining (Medlars)



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

Derivation

- Given an $m \times n$ term-by-document (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.

Minimization Challenges and Formulations [Berry et al., 2007]

- **Local Minima**: Non-convexity of functional $f(W, H) = \frac{1}{2}\|X - WH\|_F^2$ in both W and H.
- **Non-unique Solutions**: $WDD^{-1}H$ is nonnegative for any nonnegative (and invertible) D.
- **NNMF Formulations**:
  - Information theoretic – Lee and Seung (2001)
  - Compensate for feature redundancy – Guillamet et al. (2001)
  - Extract spatially localized features – Wang et al. (2004)
  - Alternative cost functions – Hamza and Brady (2006), Dhillon and Sra (2005), Cichocki et al. (2006)

Multiplicative Method (MM)

- Multiplicative update rules for W and H (Lee and Seung, 1999):
  1. Initialize W and H with nonnegative values, and scale the columns of W to unit norm.
  2. Iterate for each c, j, and i until convergence or after k iterations:
     1. $H_{cj} \leftarrow H_{cj} \frac{(W^T X)_{cj}}{(W^T WH)_{cj} + \epsilon}$
     2. $W_{ic} \leftarrow W_{ic} \frac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$
     3. Scale the columns of W to unit norm.
- Setting $\epsilon = 10^{-9}$ will suffice to avoid division by zero.

Multiplicative Method (MM) contd.

MULTIPLICATIVE UPDATE MATLAB®CODE FOR NNMF

```
W = rand(m,k);     % W initially random
H = rand(k,n);     % H initially random
for i = 1 : maxiter
    H = H .* (W^T A) ./ (W^T WH + ϵ);
    W = W .* (AH^T) ./ (WHH^T + ϵ);
end
```


Lee and Seung MM Convergence

- **Convergence**: when the MM algorithm converges to a limit point in the interior of the feasible region, the point is a *stationary point*. The stationary point **may or may not be a local minimum**. If the limit point lies on the boundary of the feasible region, one cannot determine its stationarity [Berry et al., 2007].
- **Modifications**: Gonzalez and Zhang (2005) accelerated convergence somewhat but stationarity issue remains; Lin (2005) modified the algorithm to guarantee convergence to a stationary point; Dhillon and Sra (2005) derived update rules that incorporate weights for the importance of certain features of the approximation.

Alternating Least Squares Formulation

Basic ALS Approach:
ALS algorithms exploit the convexity of W or H (not both) in the underlying optimization problem. The basic iteration involves

(LS) Solve for H in $W^T WH = W^T X$.
(NN) Set negative elements of H to 0.
(LS) Solve for W in $HH^T W^T = HX^T$.
(NN) Set negative elements of W to 0.

ALS Recovery and Constraints:

- Unlike the MM algorithm, an element of W (or H) that becomes 0 does not have to remain 0; method can escape/recover from a *poor* path.
- Paatero (1999) and Langville et al. (2006) have improved the computational complexity of the ALS approach; sparsity and nonnegativity contraints are enforced.

Alternating Least Squares Algorithms, contd.

ALS Convergence:

- Polak (1971) showed that every limit point of a sequence of alternating variable iterates is a stationary point.
- Lawson and Hanson (1995) produced the Non-Negative Least Squares (NNLS) that was shown to converge to a local minimum.
- The price for convergence of ALS algorithms is the usual high cost per iteration – Bro and de Jong (1997).

Enforcing Statistical Sparsity (Hoyer)

- 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) [Berry and Browne, 2005].

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 (proof still needed).

GD-CLS Algorithm

1. Initialize W and H with nonnegative values, and scale the columns of W to unit norm.
2. Iterate until convergence or after k iterations:
   1. $W_{ic} \leftarrow W_{ic} \dfrac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$, for c and i
   2. Rescale the columns of W to unit norm.
   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.

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 NNMF (CNMF) iteration:

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

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

Improving Feature Interpretability

Gauging Parameters for Constrained Optimization

How sparse (or smooth) should factors (W, H) be to produce as many interpretable features as possible?

To what extent do different norms (L_1, L_2, L_∞) improve/degradate feature quality or span? At what cost?

Can a common nonnegative feature space be built from objects in both images and text? Are there opportunities for multimodal document similarity?

Anomaly Classification (ASRS)

- Classify events described by documents from the Airline Safety Reporting System (ASRS) into 22 anomaly categories; contest from SDM07 Text Mining Workshop.
- General Text Parsing (GTP) Software Environment in C++ [Giles et al., 2003] used to parse both ASRS training set and a combined ASRS training and test set:

Dataset	Terms	ASRS Documents
Training	15,722	21,519
Training+Test	17,994	28,596 (7,077)

- Web links:
  GTP: http://www.cs.utk.edu/∼lsi
  ASRS: http://www.cs.utk.edu/tmw07

Term Weighting Schemes

- **Assessment of Term Importance**: 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 $d_j = 1$).
- **Common Term Weighting Choices**:

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

Parameterization of NNMF Classifier

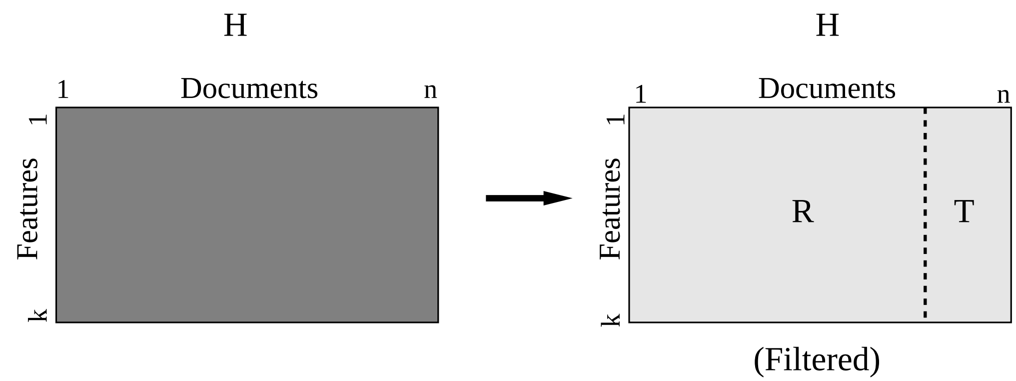
- Important Control Parameters:

  α, the threshold on the relevance score or (target value) t_{ij} for document i and anomaly/label j; we use **R** submatrix of **H** to cluster documents by the k features — assume documents describing similar anomalies share similar features.

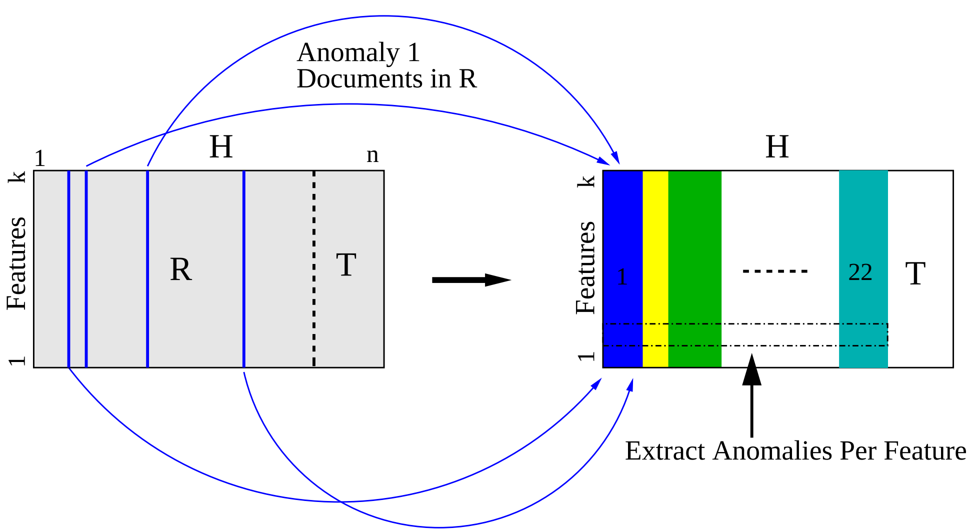
  δ, the threshold on the column elements of **H**, which will filter out the association of features with both the training (**R**) and test (**T**) documents;

  σ, the percentage of documents used to define the training set (or number of columns of **R**).

Initialization Schematic



Anomaly to Feature Mapping and Scoring Schematic

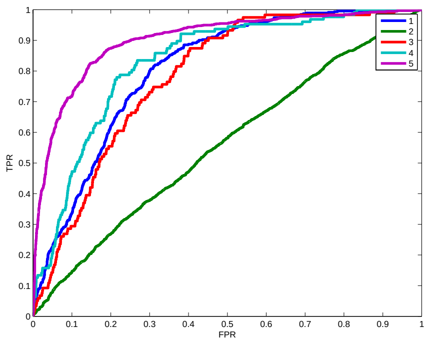


Training/Testing Performance (ROC Curves)

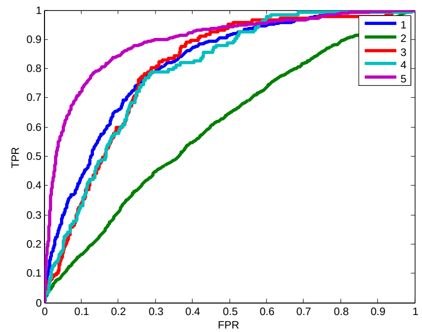
- Best/Worst ROC curves (False Positive Rate versus True Positive Rate)

Anomaly	**Type (Description)**	**ROC Area** Training	**ROC Area** Contest
22	Security Concern/Threat	.9040	.8925
5	Incursion (collision hazard)	.8977	.8716
4	Excursion (loss of control)	.8296	**.7159**
21	Illness/Injury Event	.8201	.8172
12	Traffic Proximity Event	.7954	.7751
7	Altitude Deviation	.7931	.8085
18	Aircraft Damage/Encounter	.7250	.7261
11	Terrain Proximity Event	.7234	**.7575**
9	Speed Deviation	.7060	.6893
10	Uncommanded (loss of control)	.6784	.6504
13	Weather Issue	.6287	.6018
2	Noncompliance (policy/proc.)	.6009	**.5551**

ROC Curves for Anomalies 1–5 (Test/Training)

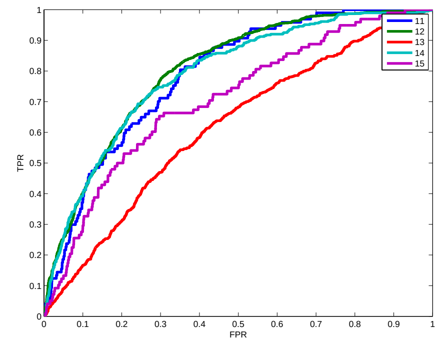


Training

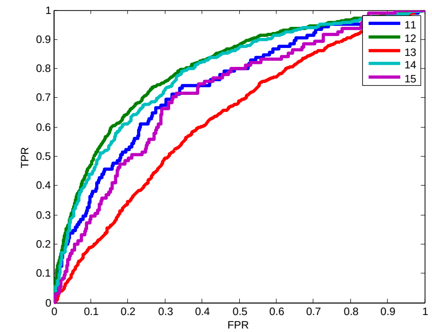


Contest

ROC Curves for Anomalies 11–15 (Test/Training)

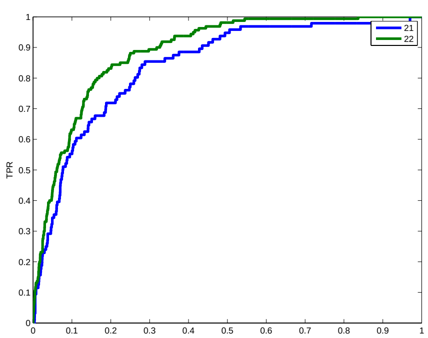


Training

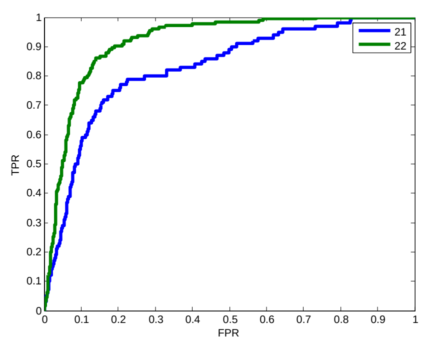


Contest

ROC Curves for Anomalies 21, 22 (Test/Training)



Training



Contest

Effects of Alternative Norms in the Objective Function

- CVX: Matlab Software for Disciplined Convex Programming (Grant, Boyd, and Ye – Stanford) used for ALP (Alternate Linear Programming)
- Factor first 100 ASRS docs only (X is 773×100) with $k = 10, 20, 40$ features (rank)

	Area(ROC) for Rank-10			Anomaly Count in	
Anomaly	L_1	L_2	L_∞	Training Set	Test Set
2	**0.6124**	0.3971	**0.6986**	46	11
5	**0.6914**	0.6173	0.2099	5	3
7	0.4560	0.5600	0.5360	11	5
8	**0.7037**	0.3580	**0.5062**	9	3
10	0.6071	0.6071	0.3929	1	2
12	**0.5900**	0.5650	0.3900	10	10
13	**0.7500**	0.6964	0.6607	7	2
18	**0.8750**	0.4286	**0.6429**	2	2

Effects of Alternative Norms in the Objective Function

Anomaly	Area(ROC) for Rank-20 L_1	Area(ROC) for Rank-20 L_2	Area(ROC) for Rank-20 L_∞	Area(ROC) for Rank-40 L_1	Area(ROC) for Rank-40 L_2	Area(ROC) for Rank-40 L_∞
2	0.4067	0.5981	0.3062	**0.5215**	0.4593	**0.5550**
5	0.3951	0.6543	0.5679	0.3704	**0.7037**	0.6296
7	0.5200	0.5680	0.4160	0.3040	**0.6320**	0.4080
8	0.6296	**0.7160**	0.5556	0.4815	0.5062	0.3951
10	**0.6786**	0.5357	**0.6429**	0.6429	0.5714	**0.6786**
12	0.5200	0.6950	0.5000	0.6800	**0.7600**	0.6400
13	**0.4464**	0.4286	**0.6429**	0.6071	**1.0000**	0.5179
18	**0.6786**	0.2679	**0.6607**	0.8036	**0.9107**	0.6964

Effects of Alternative Norms in the Objective Function

- Cost of computing a rank-40 NMF on a Dell OptiPlex 745 (2.4-GHz dual processor, 4-MB cache, 2-GB RAM); 10 repetitions with same (initial) random W, H.

	Computation of $X = WH$ L_1	L_2	L_∞
Mean (secs)	1416.255	**0.047**	2236.850
Mean (mins)	23.604	**0.001**	37.281
Variance	81.124	0	163.147
StdDev	9.007	0.002	12.773

Effects of Alternative Norms in the Objective Function

- Sparsity of NMF factors (without constraints) can vary with L_p norm (and rank); $X^{773\times 100}$

Rank	Factor	Number of Nonzeros L_1	L_2	L_∞
10	W	7730	7730	7730
10	H	**728**	1000	1000
20	W	15140	15460	15460
20	H	**1182**	2000	2000
40	W	27831	30920	30920
40	H	**70**	4000	4000

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 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.

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 [Berry and Browne, 2005].
- 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

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

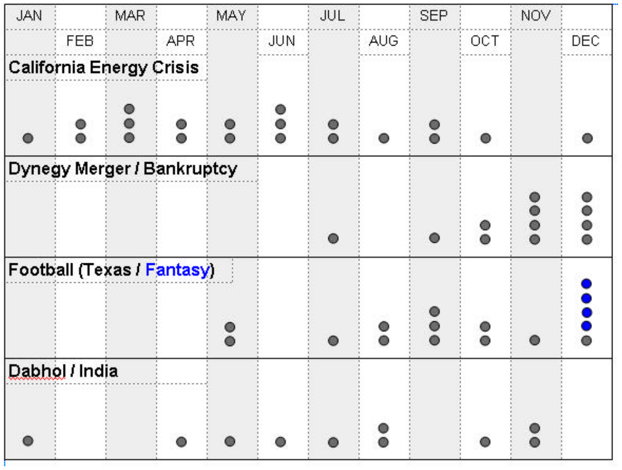
(Cluster size $\equiv$ no. of H^k elements $>$ $row_{max}/10$)

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, $\lambda = 0.1$

(*New York Times*, May 22, 2005)

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

Annotation Project

- Subset of 2001 PRIVATE collection:

Month	Total	Classified	Usable
Jan,Sep	5591	1100	699
Feb	2804	900	460
Mar	3525	1200	533
Apr	4273	1500	705
May	4261	1800	894
June	4324	1025	538
Total	24778	7525	3829

- Approx. 40 topics identified after NNMF initial clustering with $k = 50$ features.

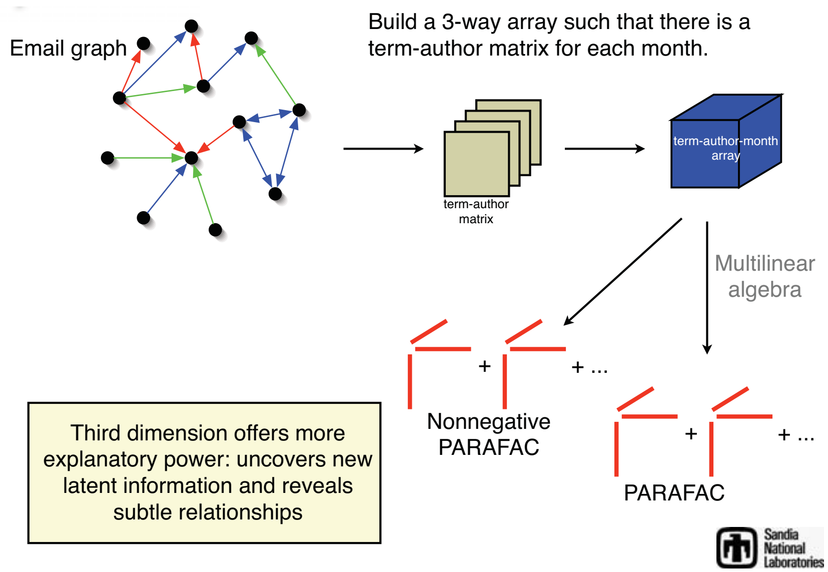
Annotation Project, contd.

- Human classfiers: M. Browne (extensive background reading on Enron collapse) and B. Singer (junior Economics major).
- Classify email content versus type (see UC Berkeley Enron Email Analysis Group `http://bailando.sims.berkeley.edu/enron_email.html`
- As of June 2007, distributed by the by U. Penn LDC (Linguistic Data Consortium); see `www.ldc.upenn.edu`

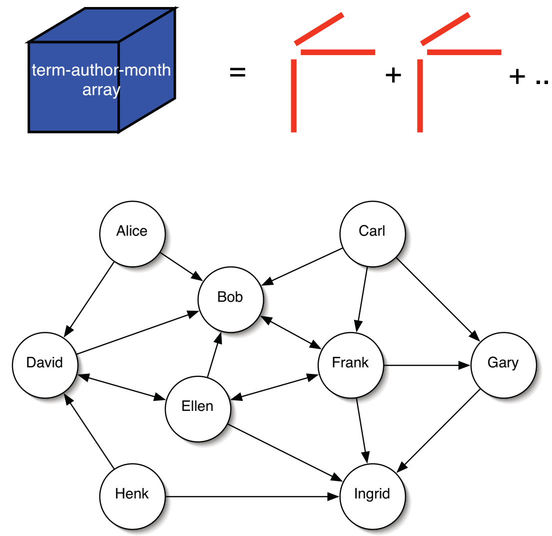
Citation:

Dr. Michael W. Berry, Murray Browne
and Ben Signer, 2007
2001 Topic Annotated Enron Email Data Set
Linguistic Data Consortium, Philadelphia

Multidimensional Data Analysis via PARAFAC



Temporal Assessment via PARAFAC



Mathematical Notation

- Kronecker product

$$A \otimes B = \begin{pmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{pmatrix}$$

- Khatri-Rao product (columnwise Kronecker)

$$A \odot B = \begin{pmatrix} A_1 \otimes B_1 & \cdots & A_n \otimes B_n \end{pmatrix}$$

- Outer product

$$A_1 \circ B_1 = \begin{pmatrix} A_{11}B_{11} & \cdots & A_{11}B_{m1} \\ \vdots & \ddots & \vdots \\ A_{m1}B_{11} & \cdots & A_{m1}B_{m1} \end{pmatrix}$$

PARAFAC Representations

- PARAllel FACtors (Harshman, 1970)
- Also known as CANDECOMP (Carroll & Chang, 1970)
- Typically solved by Alternating Least Squares (ALS)

Alternative PARAFAC formulations

$$X_{ijk} \approx \sum_{i=1}^{r} A_{ir} B_{jr} C_{kr}$$

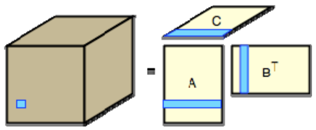
$$\mathcal{X} \approx \sum_{i=1}^{r} A_i \circ B_i \circ C_i, \text{ where } \mathcal{X} \text{ is a 3-way array (tensor).}$$

$$X_k \approx A \,\mathrm{diag}(C_{k:})\, B^T, \text{ where } X_k \text{ is a tensor slice.}$$

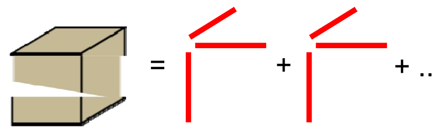
$$X^{I \times JK} \approx A(C \odot B)^T, \text{ where } X \text{ is matricized.}$$

PARAFAC (Visual) Representations

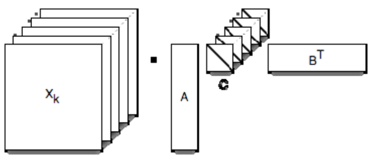
Scalar form



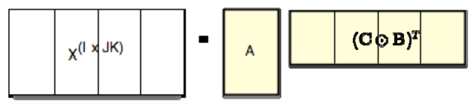
Outer product form



Tensor slice form



Matrix form



Nonnegative PARAFAC Algorithm

- Adapted from (Mørup, 2005) and based on NNMF by (Lee and Seung, 2001)

$$\begin{aligned} ||X^{I\times JK} - A(C\odot B)^T||_F &= ||X^{J\times IK} - B(C\odot A)^T||_F \\ &= ||X^{K\times IJ} - C(B\odot A)^T||_F \end{aligned}$$

- Minimize over A, B, C using multiplicative update rule:

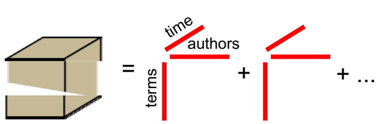
$$A_{i\rho} \leftarrow A_{i\rho}\frac{(X^{I\times JK}Z)_{i\rho}}{(AZ^TZ)_{i\rho}+\epsilon}, \quad Z=(C\odot B)$$

$$B_{j\rho} \leftarrow B_{j\rho}\frac{(X^{J\times IK}Z)_{j\rho}}{(BZ^TZ)_{j\rho}+\epsilon}, \quad Z=(C\odot A)$$

$$C_{k\rho} \leftarrow C_{k\rho}\frac{(X^{K\times IJ}Z)_{k\rho}}{(CZ^TZ)_{k\rho}+\epsilon}, \quad Z=(B\odot A)$$

Discussion Tracking Using Year 2001 Subset

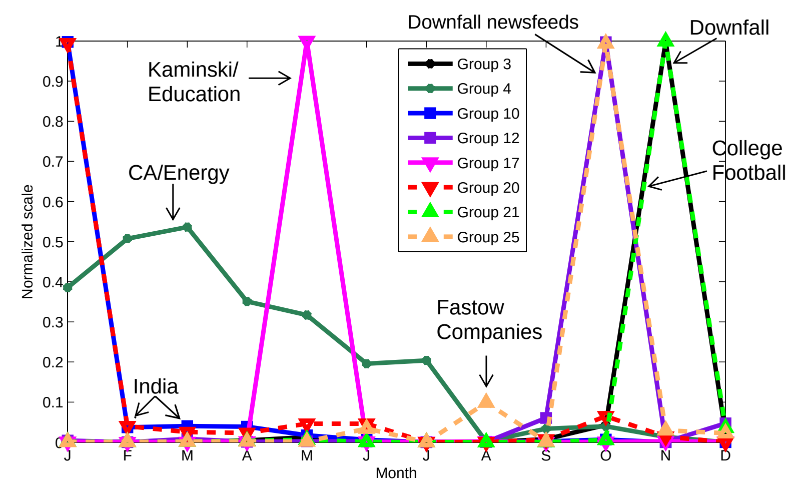
- 197 authors (`From:user_id@enron.com`) monitored over 12 months;
- Parsing $34,427$ email subset with a base dictionary of $121,393$ terms (derived from $517,431$ emails) produced $69,157$ unique terms; (term-author-month) array X has ~ 1 million nonzeros.
- Rank-25 tensor: A $(69,157\times 25)$, B (197×25), C (12×25)



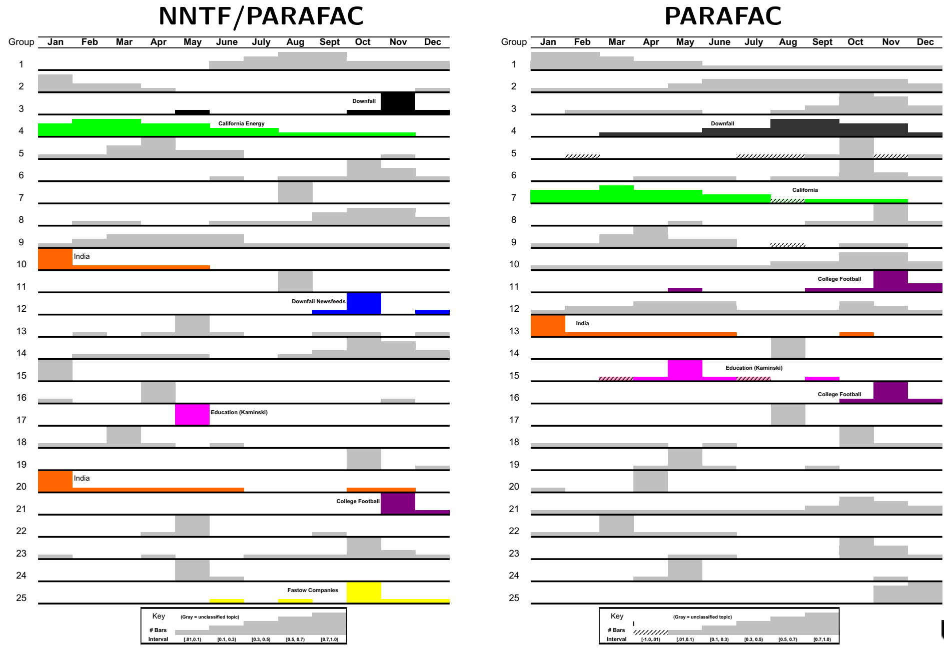
Month	Emails	Month	Emails
Jan	7,050	Jul	2,166
Feb	6,387	Aug	2,074
Mar	6,871	Sep	2,192
Apr	7,382	Oct	5,719
May	5,989	Nov	4,011
Jun	2,510	Dec	1,382

Tensor-Generated Group Discussions

- NNTF Group Discussions in 2001
- 197 authors; 8 distinguishable discussions
- "Kaminski/Education" topic previously unseen

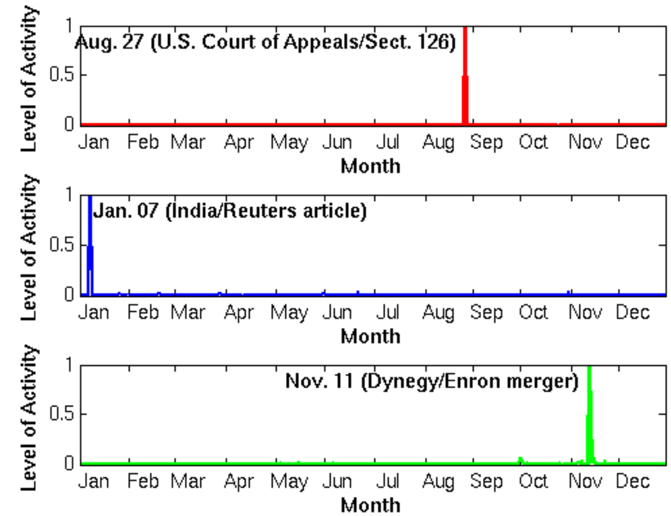


Gantt Charts from PARAFAC Models



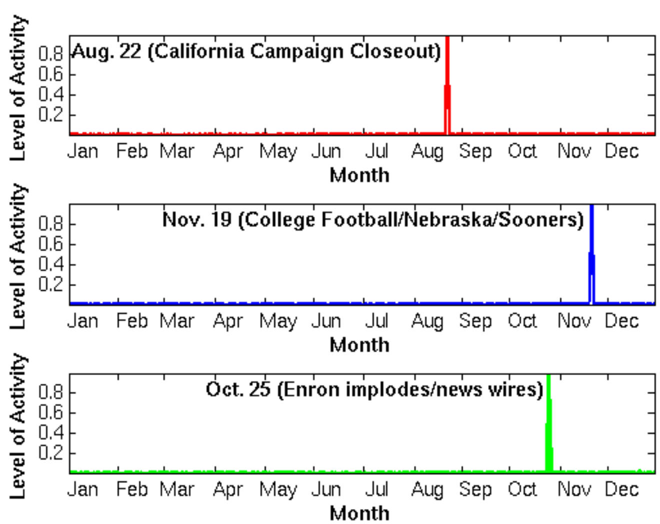
Day-level Analysis for PARAFAC (Three Groups)

- Rank-25 tensor for 357 out of 365 days of 2001: A $(69,157 \times 25)$, B (197×25), C (357×25)
- Groups 3,4,5 (out of 25 from C) [Bader et al., 2008]:



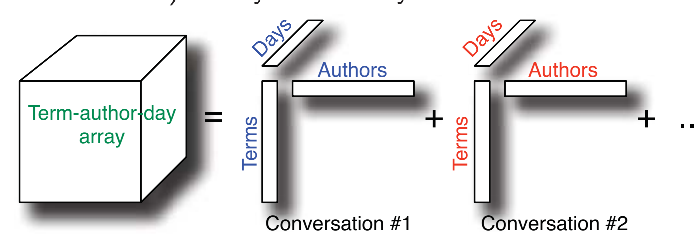
Day-level Analysis for NN-PARAFAC (Three Groups)

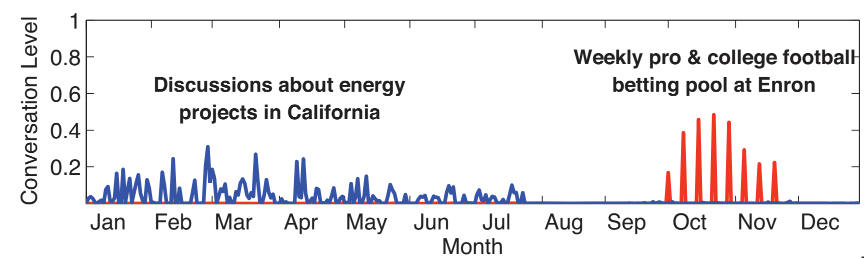
- Rank-25 tensor (best minimizer) for 357 out of 365 days of 2001: A $(69,157 \times 25)$, B (197×25), C (357×25)
- Groups 1,7,8 (out of 25 from C):



Day-level Analysis for NN-PARAFAC (Two Groups)

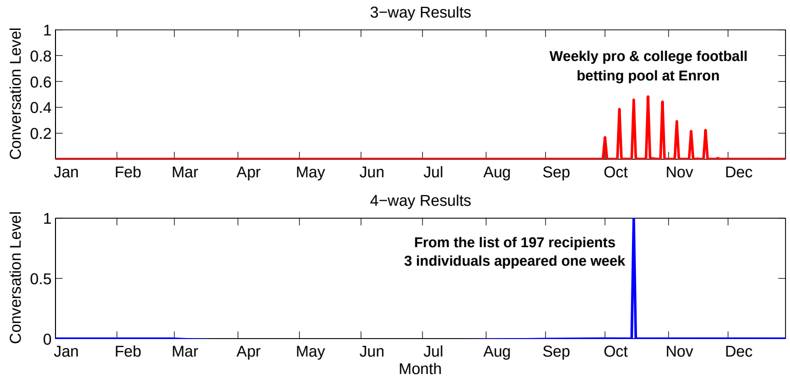
- Groups 20 (California Energy) and 9 (Football) (from C factor of best minimizer) in day-level analysis of 2001:





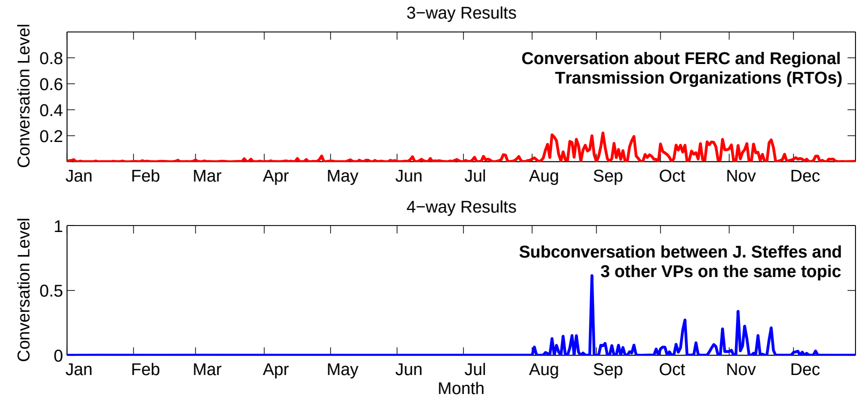
Four-way Tensor Results (Sept. 2007)

- Apply NN-PARAFAC to term-author-recipient-day array ($39,573 \times 197 \times 197 \times 357$); construct a rank-25 tensor (best minimizer among 10 runs).
- Goal: track more focused discussions between individuals/ small groups; for example, betting pool (football).



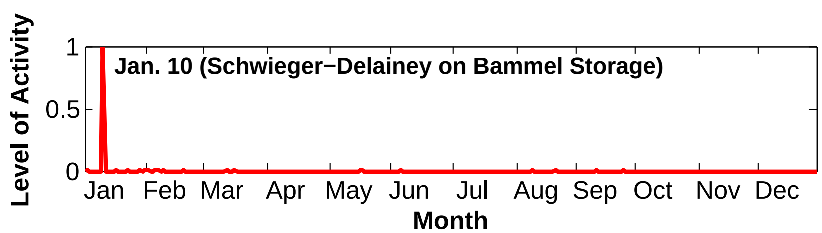
Four-way Tensor Results (Sept. 2007)

- Four-way tensor may track subconversation already found by three-way tensor; for example, RTO (Regional Transmission Organization) discussions.



Four-way Tensor Results (October 2007)

- Four-way tensor exposed conversation confirming bank fraud related to the natural gas reserves in the Bammel Storage field (Texas)—"The Enron whistle-blower who wasn't" by G. Farrell, **USA Today**, Oct. 11, 2007



NNTF Optimal Rank?

- No known algorithm for computing the rank of a k-way array for $k \geq 3$ [Kruskal, 1989].
- The maximum rank is **not a closed set** for a given random tensor.
- The maximum rank of a $m \times n \times k$ tensor is unknown; one weak inequality is given by

$$\max\{m, n, k\} \leq \text{ rank } \leq \min\{m \times n, m \times k, n \times k\}$$

- For our rank-25 NNTF, the size of the relative residual norm suggests we are still far from the maximum rank of the 3-way and 4-way arrays.

Conclusions (NNMF for ASRS)

- Training phase was a good predictor of performance (for most anomaly categories).
- Obvious room for improvement in matching certain anomalies (e.g., 2. Noncompliance).
- Summarization of anomalies using NNMF features needs further work.
- Effects of sparsity contraints on NNMF versus element-wise filtering of **H** should be studied.
- Effects of using different L_p norms in the objective function need more analysis.

Conclusions (NNMF/NNTF for Enron)

- GD-CLS/NNMF 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 (log-entropy) terms; yields potential **control vocabulary** for topic tracking.
- Surveillance systems based on NNMF/NNTF algorithms show promise for monitoring discussions without the need to isolate or perhaps incriminate individuals.
- Potential applications include the monitoring/tracking of company morale, employee feedback to policy decisions, extracurricular activities, and blog discussions.

Research Tracks

- Further work needed in determining effects of alternative term weighting schemes (for X) and choices of control parameters (e.g., α, β) for CNMF and NNTF/PARAFAC.
- How many dimensions (factors) for NNTF/PARAFAC are really needed for mining electronic mail and similar corpora? And, at what **scale** should each dimension be measured (e.g., **time**)?
- Compare convergence and accuracy of CNMF with NMU (Nonnegative Matrix Underapproximation); joint work with N. Gillis and F. Glineur (Univ. of Louvain, Belgium)

Improving Summarization and Steering

What versus why:

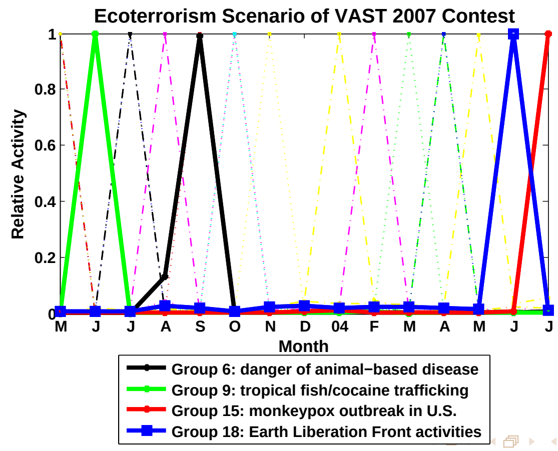
Extraction of textual concepts still requires human interpretation (in the absence of ontologies or domain-specific classifications).

How can previous knowledge or experience be captured for feature matching (or pruning)?

To what extent can feature vectors be annotated for future use or as the text collection is updated? What is the cost for updating NNMF/NNTF models?

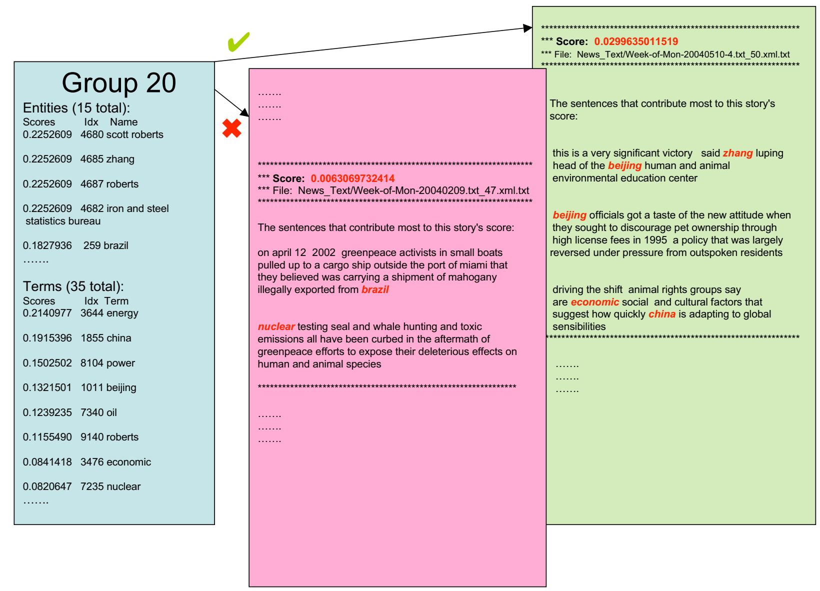
NNTF for Visual Analytics (VA)

- VAST 2007 Contest: 1,455 news stories/emails/blog entries with underlying ecoterrorism activity to be uncovered.
- Who/What/When/Where questions using tagged entities (Person, Location, Organization, Money) and context (terms). (See http://www.cs.umd.edu/hcil/VASTcontest07)



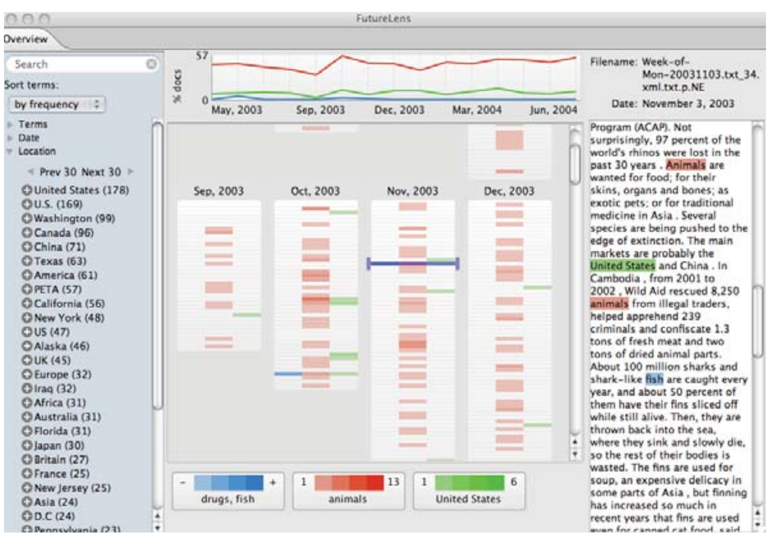
NNTF for Visual Analytics (VA)

- Score documents (news stories) against terms and entities from identifiable (classifiable) NTF factors:



FutureLens: NNTF Output Visualization

- Java-based visualization environment (adapted from **FeatureLens** at HCIL, Univ. of Maryland)



For Further Reading

- B.W. Bader, M.W. Berry, and M. Browne.
  Discussion Tracking in Enron Email Using PARAFAC.
  in *Survey of Text Mining II: Clustering, Classification, and Retrieval*, M.W. Berry and M. Castellanos (Eds.), Springer-Verlag, London, 2008:147-163.
- M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons.
  Alg. and Applic. for Approx. Nonnegative Matrix Factorization.
  Comput. Stat. & Data Anal. 52(1):155-173, 2007.
- F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons.
  Document Clustering Using Nonnegative Matrix Factorization.
  Info. Proc. & Management 42(2):373-386, 2006.
- M.W. Berry and M. Browne.
  Email Surveillance Using Nonnegative Matrix Factorization.
  Comp. & Math. Org. Theory 11:249-264, 2005.

For Further Reading (contd.)

- P. Hoyer.
  Nonnegative Matrix Factorization with Sparseness Constraints.
  J. Machine Learning Research 5:1457-1469, 2004.
- J.T. Giles and L. Wo and M.W. Berry.
  GTP (General Text Parser) Software for Text Mining.
  Software for Text Mining, in Statistical Data Mining and Knowledge Discovery. CRC Press, Boca Raton, FL, 2003:455-471.
- W. Xu, X. Liu, and Y. Gong.
  Document-Clustering based on Nonneg. Matrix Factorization.
  Proceedings of SIGIR'03, Toronto, CA, 2003:267-273.
- J.B. Kruskal.
  Rank, Decomp. , and Uniqueness for 3-way and n-way Arrays.
  In *Multiway Data Analysis*, Eds. R. Coppi and S. Bolaso, Elsevier 1989:7-18.