Automating the Detection of Anomalies and Trends from Text

NGDM'07 Workshop
Baltimore, MD

Michael W. Berry

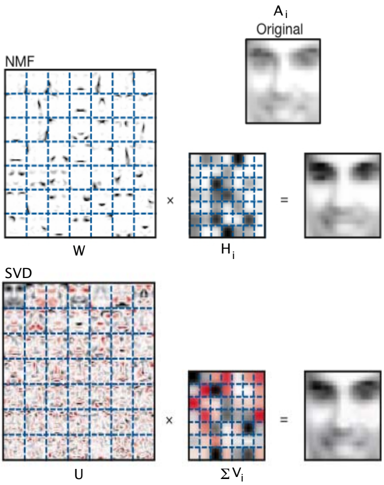
Department of Electrical Engineering & Computer Science
University of Tennessee

October 11, 2007

NNMF Origins

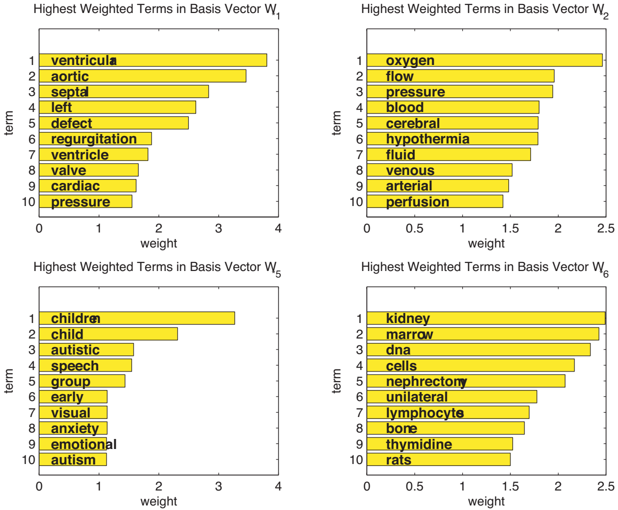
- NNMF (Nonnegative Matrix Factorization) can be used to approximate high-dimensional data having nonnegative components.
- Lee and Seung (1999) demonstrated its use as a *sum-by-parts* representation of image data in order to both identify and classify image *features*.
- Xu et al. (2003) demonstrated 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 Analogue 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.
- **Many NNMF Formulations**:
  - Lee and Seung (2001) – information theoretic formulation based on Kullback-Leibler divergence of X from WH.
  - Guillamet, Bressan, and Vitria (2001) – diagonal weight matrix Q used ($XQ \approx WHQ$) to compensate for feature redundancy (columns of W).
  - Wang, Jiar, Hu, and Turk (2004) – constraint-based formulation using Fisher linear discriminant analysis to improve extraction of spatially localized features.
  - Other Cost Function Formulations – Hamza and Brady (2006), Dhillon and Sra (2005), Cichocki, Zdunek, and Amari (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. $H_{cj} \leftarrow H_{cj} \dfrac{(W^T X)_{cj}}{(W^T WH)_{cj} + \epsilon}$
     2. $W_{ic} \leftarrow W_{ic} \dfrac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$
     3. Scale the columns of W to unit norm.
- Setting $\epsilon = 10^{-9}$ will suffice to avoid division by zero [Shahnaz et al., 2006].

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].
- **Several modifications have been proposed**: 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.

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 (or smoothness) – 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 nonnegative feature space be built from objects in both images and text? Are there opportunities for multimodal document similarity?

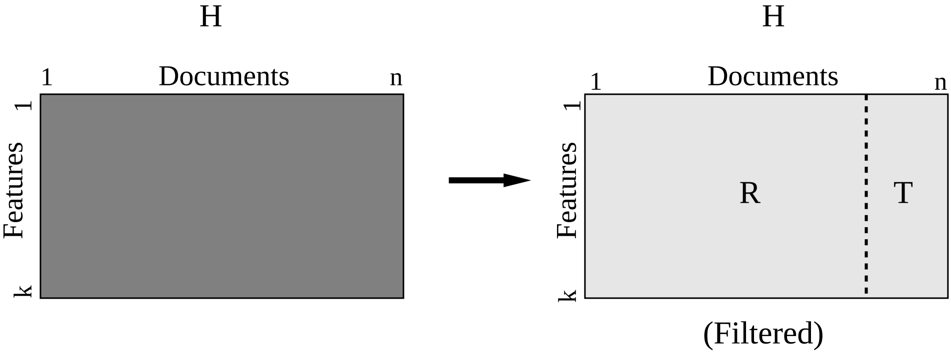
Anomaly Detection (ASRS)

- 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:

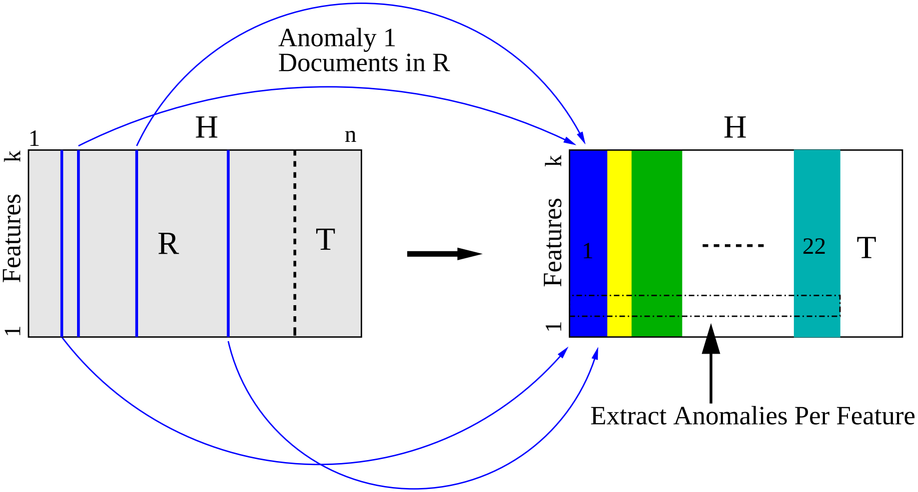
Dataset	Terms	ASRS Documents
Training	15,722	21,519
Training+Test	17,994	28,596 (7,077)

- Global and document frequency of required to be at least 2; stoplist of 493 common words used; char length of any term $\in [2, 200]$.
- Download Information:
  GTP: http://www.cs.utk.edu/∼lsi
  ASRS: http://www.cs.utk.edu/tmw07

Initialization Schematic



Anomaly to Feature Mapping and Scoring Schematic

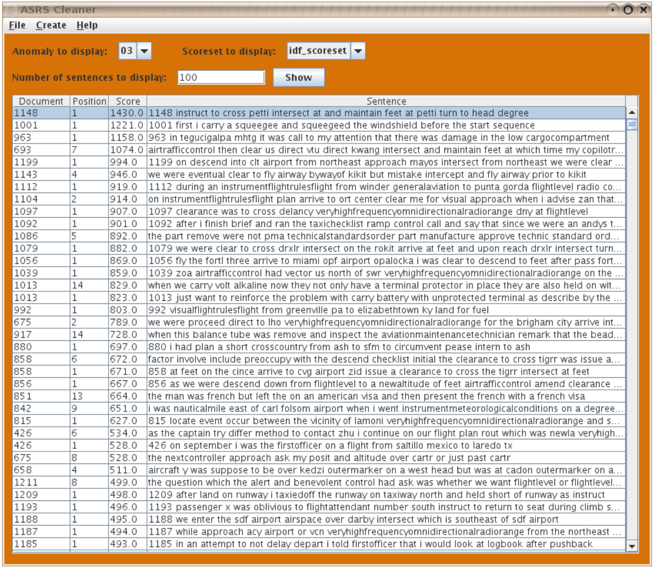


Training/Testing Performance (ROC Curves)

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

Anomaly	Type (Description)	ROC Area	
		Training	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**

Anomaly Summarization Prototype - Sentence Ranking



Document	Position	Score	Sentence
1148	1	1430.0	1148 instruct to cross petti intersect at and maintain feet at petti turn to head degree
1001	1	1221.0	1001 first i carry a squeegee and squeegeed the windshield before the start sequence
963	1	1158.0	963 in tegucigalpa mhtg it was call to my attention that there was damage in the low cargocompartment
693	7	1074.0	airtrafficcontrol then clear us direct vtu direct kwang intersect and maintain feet at which time my copilotr...
1199	1	994.0	1199 on descend into clt airport from northeast approach mayos intersect from northeast we were clear ...
1143	4	946.0	we were eventual clear to fly airway bywayof kikit but mistake intercept and fly airway prior to kikit
1112	1	919.0	1112 during an instrumentflightrulesflight from winder generalaviation to punta gorda flightlevel radio co...
1104	2	914.0	on instrumentflightrulesflight plan arrive to ort center clear me for visual approach when i advise zan that...
1097	1	907.0	1097 clearance was to cross delancy veryhighfrequencyomnidirectionalradiorange dny at flightlevel
1092	1	901.0	1092 after i finish brief and ran the taxichecklist ramp control call and say that since we were an andys t...
1086	5	892.0	the part remove were not pma technicalstandardsorder part manufacture approve technic standard ord...
1079	1	882.0	1079 we were clear to cross drxlr intersect on the rokit arrive at feet and upon reach drxlr intersect turn...
1056	1	869.0	1056 fly the fortl three arrive to miami opf airport opalocka i was clear to descend to feet after pass fort...
1039	1	859.0	1039 zoa airtrafficcontrol had vector us north of swr veryhighfrequencyomnidirectionalradiorange on the ...
1013	14	829.0	when we carry volt alkaline now they not only have a terminal protector in place they are also held on wit...
1013	1	823.0	1013 just want to reinforce the problem with carry battery with unprotected terminal as describe by the ...
992	1	803.0	992 visualflightrulesflight from greenville pa to elizabethtown ky land for fuel
675	2	789.0	we were proceed direct to lho veryhighfrequencyomnidirectionalradiorange for the brigham city arrive int...
917	14	728.0	when this balance tube was remove and inspect the aviationmaintenancetechnician remark that the bead...
880	1	697.0	880 i had plan a short crosscountry from ash to sfm to circumvent pease intern to ash
858	6	672.0	factor involve include preoccupy with the descend checklist initial the clearance to cross tigrr was issue a...
858	1	671.0	858 at feet on the cince arrive to cvg airport zid issue a clearance to cross the tigrr intersect at feet
856	1	667.0	856 as we were descend down from flightlevel to a newaltitude of feet airtrafficcontrol amend clearance ...
851	13	664.0	the man was french but left the on an american visa and then present the french with a french visa
842	9	651.0	i was nauticalmile east of carl folsom airport when i went instrumentmeteorologicalconditions on a degree...
815	1	627.0	815 locate event occur between the vicinity of lamoni veryhighfrequencyomnidirectionalradiorange and s...
426	6	534.0	as the captain try differ method to contact zhu i continue on our flight plan rout which was newla veryhigh...
426	1	528.0	426 on september i was the firstofficer on a flight from saltillo mexico to laredo tx
675	8	528.0	the nextcontroller approach ask my posit and altitude over cartr or just past cartr
658	4	511.0	aircraft y was suppose to be over kedzi outermarker on a west head but was at cadon outermarker on a...
1211	8	499.0	the question which the alert and benevolent control had ask was whether we want flightlevel or flightlevel...
1209	1	498.0	1209 after land on runway i taxiedoff the runway on taxiway north and held short of runway as instruct
1193	1	496.0	1193 passenger x was oblivious to flightattendant number south instruct to return to seat during climb s...
1188	1	495.0	1188 we enter the sdf airport airspace over darby intersect which is southeast of sdf airport
1187	1	494.0	1187 while approach acy airport or vcn veryhighfrequencyomnidirectionalradiorange from the northeast ...
1185	1	493.0	1185 in an attempt to not delay depart i told firstofficer that i would look at logbook after pushback

Sentence rank = f(global term weights) – B. Lamb

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 the NNMF (or similar) model?

Unresolved Modeling Issues

Parameters and dimensionality:

Further work needed in determining effects of alternative term weighting schemes (for X) and choices of control parameters (e.g., α, β for CNMF).

How does document (or object) clustering change with different ranks (or features)?

How should feature vectors from competing models (Bayesian, neural nets, etc.) be compared in both interpretability and computational cost?

Email Collection

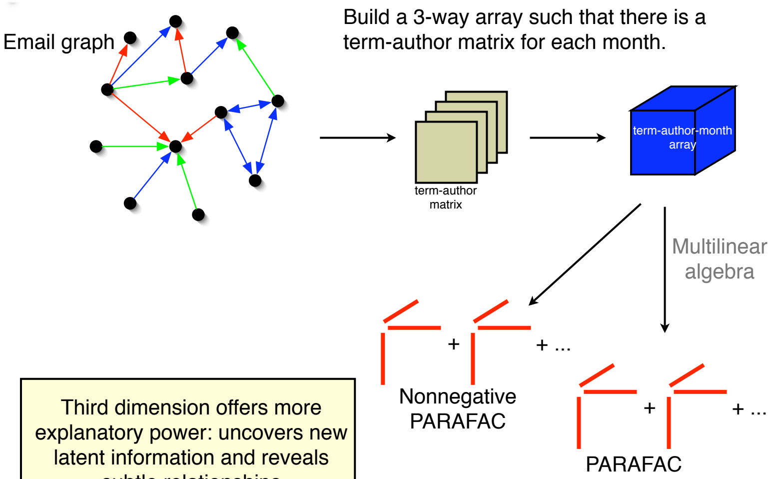
- By-product of the FERC investigation of Enron (originally contained 15 million email messages).
- This study used the improved corpus known as the Enron Email set, which was edited by Dr. William Cohen at CMU.
- This set had over 500,000 email messages. The majority were sent in the 1999 to 2001 timeframe.

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.

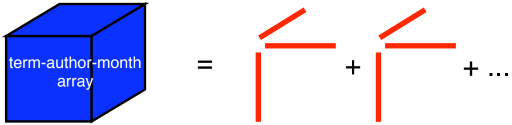
Multidimensional Data Analysis via PARAFAC

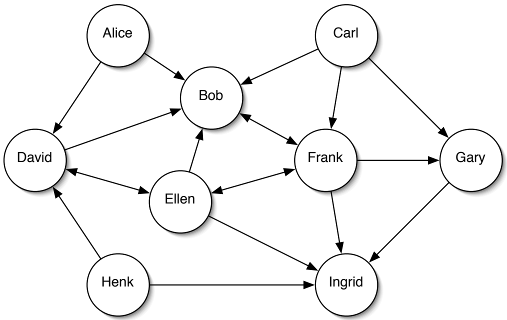
Build a 3-way array such that there is a term-author matrix for each month.



Third dimension offers more explanatory power: uncovers new latent information and reveals subtle relationships

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

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

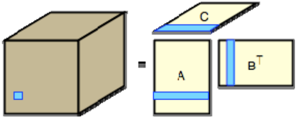
$$\mathcal{X} \approx \sum_{i=1}^{r} A_i \circ B_i \circ C_i, \text{ where } \mathcal{X} \text{ is a 3-way array (tensor).}$$

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

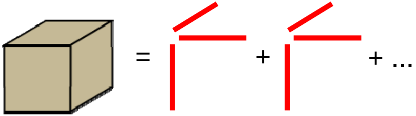
$$X^{I \times JK} \approx A(C \odot B)^T, \text{ where } X \text{ is matricized.}$$

PARAFAC (Visual) Representations

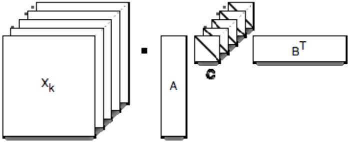
Scalar form



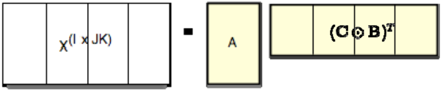
Outer product form



Tensor slice form



Matrix form



Nonnegative PARAFAC Algorithm

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

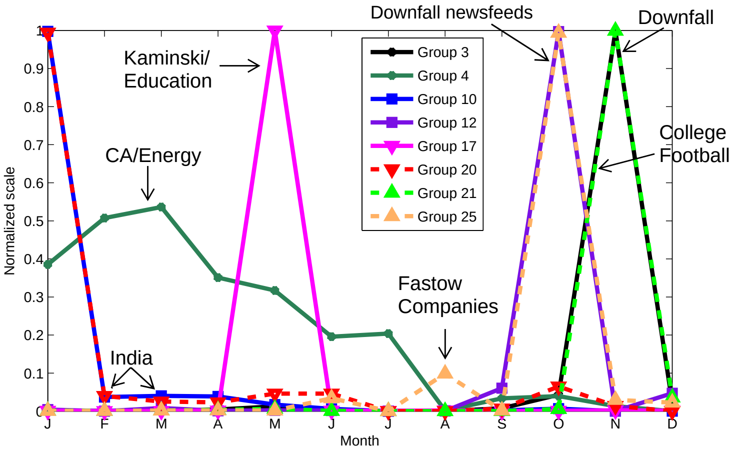
$$
\begin{aligned}
\|X^{I\times JK} - A(C\odot B)^T\|_F &= \|X^{J\times IK} - B(C\odot A)^T\|_F \\
&= \|X^{K\times IJ} - C(B\odot A)^T\|_F
\end{aligned}
$$

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

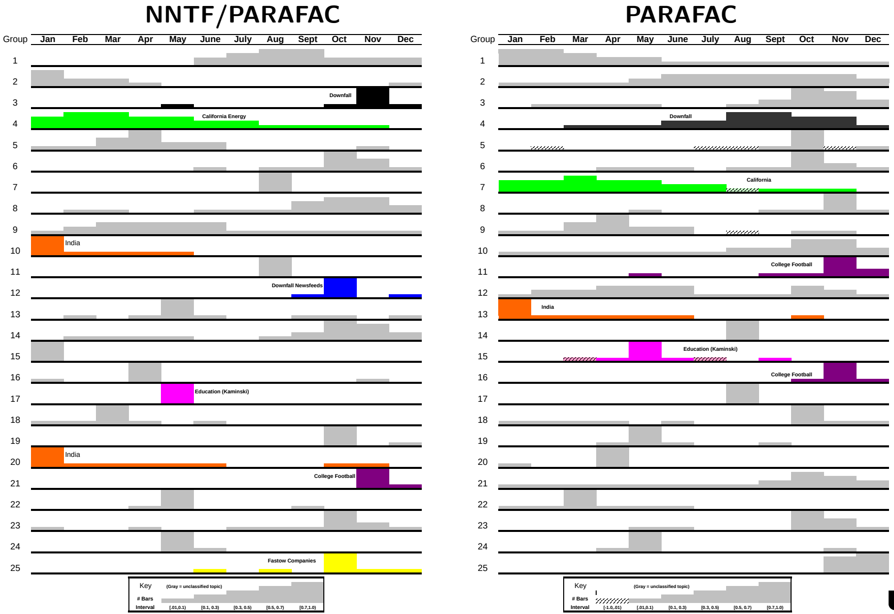
$$
\begin{aligned}
A_{i\rho} &\leftarrow A_{i\rho}\frac{(X^{I\times JK}Z)_{i\rho}}{(AZ^TZ)_{i\rho}+\epsilon}, \quad Z=(C\odot B) \\
B_{j\rho} &\leftarrow B_{j\rho}\frac{(X^{J\times IK}Z)_{j\rho}}{(BZ^TZ)_{j\rho}+\epsilon}, \quad Z=(C\odot A) \\
C_{k\rho} &\leftarrow C_{k\rho}\frac{(X^{K\times IJ}Z)_{k\rho}}{(CZ^TZ)_{k\rho}+\epsilon}, \quad Z=(B\odot A)
\end{aligned}
$$

Tensor-Generated Group Discussions

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

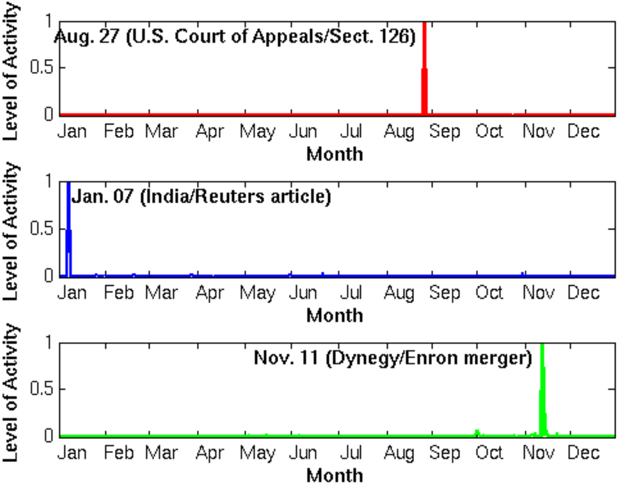


Gantt Charts from PARAFAC Models



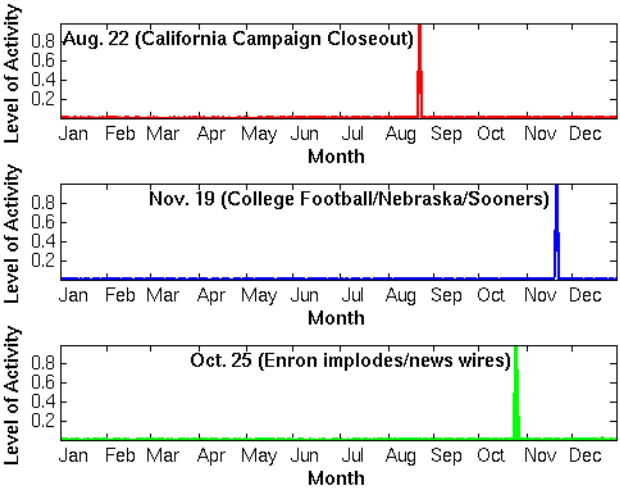
Day-level Analysis for PARAFAC (Three Groups)

- 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:



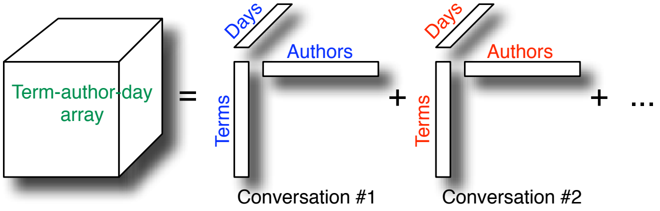
Day-level Analysis for NN-PARAFAC (Three Groups)

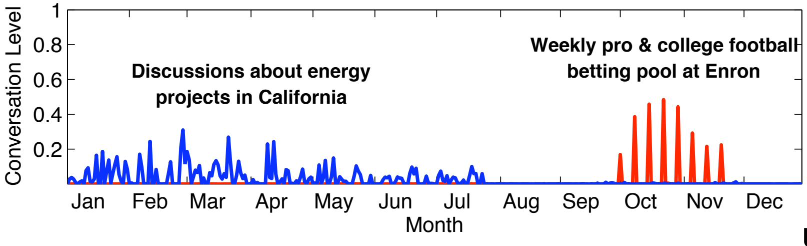
- Rank-25 tensor (best minimizer) for 357 out of 365 days of 2001: A $(69,157 \times 25)$, B (197×25), C (357×25)
- Groups 1,7,8:



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

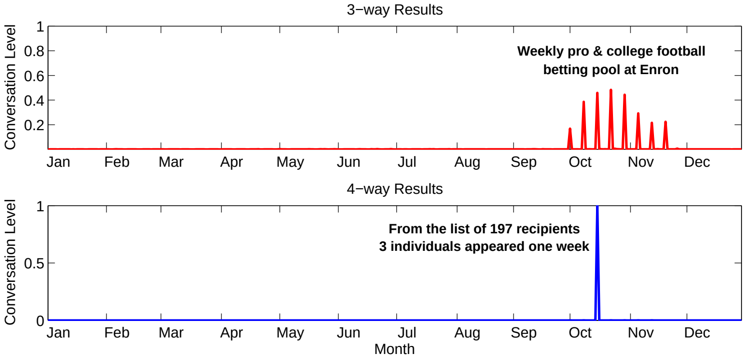
- Groups 20 (California Energy) and 9 (Football) (from C factor of best minimizer) in day-level analysis of 2001:





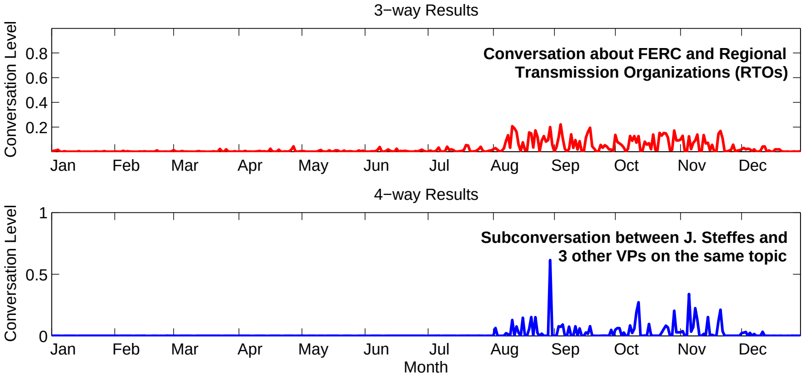
Four-way Tensor Results (Sept. 2007)

- 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.



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.

Further Reading

- 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.
- P. Hoyer.
  Nonnegative Matrix Factorization with Sparseness Constraints.
  J. Machine Learning Research 5:1457-1469, 2004.

Further Reading (contd.)

- 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, pp. 455-471.
- W. Xu, X. Liu, and Y. Gong.
  Document-Clustering based on Nonneg. Matrix Factorization.
  Proceedings of SIGIR'03, Toronto, CA, 2003, pp. 267-273.
- J.B. Kruskal.
  Rank, Decomposition, and Uniqueness for 3-way and n-way Arrays.
  In *Multiway Data Analysis*, Eds. R. Coppi and S. Bolaso, Elsevier 1989, pp. 7-18.