SAND2008-4220C

Scenario Discovery using Nonnegative Tensor Factorization *

Brett W. Bader[1], Andrey Puretskiy[2], and Michael W. Berry[2]

[1] Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-1318
bwbader@sandia.gov
[2] Department of Electrical Engineering and Computer Science, University of Tennessee, 203 Claxton Complex, Knoxville, TN 37996-3450
[puretski,berry]@eecs.utk.edu

Abstract. In the relatively new field of visual analytics there is a great need for automated approaches to both verify and discover the intentions and schemes of primary actors through time. Data mining and knowledge discovery play critical roles in facilitating the ability to extract meaningful information from large and complex textual-based (digital) collections. In this study, we develop a mathematical strategy based on nonnegative tensor factorization (NTF) to extract and sequence important activities and specific events from sources such as news articles. The ability to automatically reconstruct a plot or confirm involvement in a questionable activity is greatly facilitated by our approach. As a variant of the PARAFAC multidimensional data model, we apply our NTF algorithm to the terrorism-based scenarios of the VAST 2007 Contest dataset to demonstrate how term-by-entity associations can be used for scenario/plot discovery and evaluation.

Key words: nonnegative tensor factorization, PARAFAC, scenario discovery, VAST 2007, visual analytics

1 Introduction

Visual analytics is the science of analytical reasoning supported by the highly interactive visual interface. Tools for visual analytics are designed to synthesize information; derive insight from large, dynamic, and often conflicting data; and facilitate the mining of such data for both expected and unexpected associations that can be both verified and easily communicated. The discipline of visual analytics is very interdisciplinary and extends well beyond traditional scientific and information visualization to include statistics, mathematics, knowledge representation, management and discovery technologies, cognitive and perceptual sciences, and decision sciences.

* This research was sponsored by the United States Department of Energy and by Sandia National Laboratory, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE–AC04–94AL85000.

In this study, we use nonnegative tensor factorization (NTF) techniques to extract term-by-entity associations from textual describing the activities and events associated with a possible terrorism-based plot. As shown in [1], nonnegative tensor factorization based on the well-known PARAFAC [2] model for multidimensional data can be effective in extracting important topics of discussions from media such as electronic mail. How such tensor factorizations would fare for (automatically) exposing the major (or sub-major) plots of a scenario hidden in the details of news stories, blog entries, and similar textual data was the question we sought to address in this work. The fictious terrorist activities created by Whiting et al. for the VAST 2007 Contest [3] provides an excellent example of textual data that is needed to address that fundamental question. We will first describe the scenario provided by this dataset in Section 2 followed by a discussion of our methods of analysis using NTF in Section 3. Details of how the data was processed is provided in Section 4, and our analysis and observations using NTF are summarized in Section 5. Concluding remarks and a brief discussion of future work are given in Section 6.

2 Scenario Dataset

The VAST 2007 Contest [3], a participation category of the IEEE VAST 2007 Symposium, was designed to promote the development of benchmark data sets and metrics for visual analytics as well as to establish a forum to advance visual analytics evaluation methods. The dataset associated with the contest consisted of news stories and blog entries, along with background information and some multimedia materials (small maps and data tables). Participants in the contest were asked to investigate a major law enforcement/counter-terrorism scenario, form their hypotheses, and collect supporting evidence. Tasks that each team/entry were expected to address included *i*) processing the text and multimedia information to identify entities of interest (e.g., people, places and activities), *ii*) depicting this information visually using interactive visualizations and other tools to aid in the analysis of the information; *iii*) answering specific contest questions based on the analysis; and *iv*) producing a video demonstration of their system showing how they arrived at those answers. In our study, we demonstrate the use nonnegative tensor factorization for (automated) knowledge discovery that would facilitiate the first three of tasks mentioned above.

2.1 Scenario

The scenario depicted in the VAST 2007 Contest involved an emergency related to wildlife law enforcement occurring in the fall of 2004. Endangered species issues and ecoterrorism played key roles in the underlying terrorist activity. The heterogeneous data used to describe the situation included text, images, and some statistics. The activities of certain animal rights groups, such as the People for the Ethical Treatments of Animals (PETA) and Earth Liberation Front (ELF), are involved with the plot but are not the primary or interesting parties for investigation.

2.2 Evaluating Solutions

Entries (or answers) submitted to the VAST 2007 Contest were judged according to the correctness of the answers to the questions and the evidence provided. Points were awarded for correct answers and subtracted for incorrect answers. A more subjective assessment of the quality of the displays, interactions and support for the analytical process was also provided. Such an assessment was based on the visuals and description of the analytic process (including the video).

Participants were required to answer the questions (who, what and where) on an answer form, and for each response they were required to identify the most relevant documents or other materials from the dataset as evidence. In a *debriefing* section of the answer form, contestants must describe the plot(s) and subplots(s) and how people, motivations, activities and locations are part of the plot, their relationship, and any uncertainties or information gaps that exist. This debriefing is judged on both accuracy and clarity. A short list of the questions each entry was required to answer are provided below:

- (**Who**) Who are the players engaging in questionable activities in the plot(s)? When appropriate, specify the association they are associated with.

- (**When/What**) What events occurred during this time frame that are most relevant to the plot(s)?

- (**Where**) What locations are most relevant to the plot(s)?

In the next section, we describe how linear algebra/NTF and subsequent MATLAB software can be used as decision support tools to answer the questions above and thereby provide a more scalable and time-efficient solution to the scenario analysis needs of security experts.

3 Analysis Methods

Before presenting an overview of the algorithm and software needed to implement our nonnegative tensor factorization (NTF) model, we formerly define all the important notations associated with tensor mathematics.

3.1 Notation

We use the following notation. We denote scalars by lowercase letters (a), vectors by boldface lowercase letters ($\mathbf{a}$), matrices by boldface uppercase letters ($\mathbf{A}$), and n-way arrays or tensors by boldface script letters ($\boldsymbol{\mathcal{X}}$).

We denote the ith element of vector $\mathbf{a}$ by a_i. We denote the jth column of matrix $\mathbf{A}$ by $\mathbf{a}_j$ and element (i, j) by a_{ij}. We denote element (i, j, k) of a third-order tensor $\boldsymbol{\mathcal{X}}$ by x_{ijk}.

The symbol $\circ$ denotes the outer (tensor) product, which is defined for two column vectors as

$$\mathbf{a} \circ \mathbf{b} \equiv \mathbf{a}\mathbf{b}^T = \begin{pmatrix} a_1 b_1 & \cdots & a_1 b_m \\ \vdots & \ddots & \vdots \\ a_m b_1 & \cdots & a_m b_m \end{pmatrix}.$$

The symbol $*$ denotes the Hadamard (i.e., elementwise) matrix product,

$$\mathbf{A} * \mathbf{B} \equiv \begin{pmatrix} a_{11} b_{11} & \cdots & a_{1n} b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} b_{m1} & \cdots & a_{mn} b_{mn} \end{pmatrix}.$$

The symbol $\odot$ denotes the Khatri-Rao product (columnwise Kronecker) [4] of two matrices with the same number of columns. Given two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, the Khatri-Rao product is defined as

$$\mathbf{A} \odot \mathbf{B} \equiv \big(\mathbf{a}_1 \otimes \mathbf{b}_1 \cdots \mathbf{a}_n \otimes \mathbf{b}_n\big) = \begin{pmatrix} a_{11}\mathbf{b}_1 & \cdots & a_{1n}\mathbf{b}_n \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{b}_1 & \cdots & a_{mn}\mathbf{b}_n \end{pmatrix},$$

where the symbol $\otimes$ denotes the Kronecker product.

It is helpful in some cases to *vectorize* and/or *matricize* a tensor by simply rearranging the elements of $\mathcal{X}$ into a vector or matrix, respectively. Although different notations exist, we use the notation in [4]. For a three-way array $\mathcal{X}$ of size $m \times n \times p$, the notation $\mathbf{X}^{(m \times np)}$ represents a matrix of size $m \times np$ in which the index over n runs the fastest over the columns and p the slowest. Other permutations, such as $\mathbf{X}^{(p \times nm)}$, are possible by changing the row index and the fast/slow column indices.

The norm of a tensor, $\| \mathcal{X} \|$, is the same as the Frobenius norm of the matricized array, i.e.,

$$\| \mathcal{X} \|^2 \equiv \sum_{i,j,k} (x_{ijk})^2.$$

3.2 PARAFAC and the Nonnegative Tensor Factorization

In 1970, Harshman [2] proposed a tensor decomposition called Parallel Factors (PARAFAC), which is also known as Canonical Decomposition (CANDECOMP), as developed by Carroll and Chang [5] in the same year. For a comprehensive review of tensor decompositions and their applications, see Kolda and Bader [6].

Given a third-order tensor $\mathcal{X}$ of size $m \times n \times p$ and a desired approximation rank r, the PARAFAC model approximates $\mathcal{X}$ as a sum of r rank-1 tensors formed by the outer product of three vectors. It is convenient to group each set of r vectors together as matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, which we call factor matrices. We may express the PARAFAC model in a number of ways using different notations. Here are three equivalent representations for a third-order tensor $\mathcal{X}$:

$$\begin{aligned} x_{ijk} &\approx \sum_{l=1}^{r} a_{il} b_{jl} c_{kl}, \\ \mathcal{X} &\approx \sum_{l=1}^{r} \mathbf{a}_l \circ \mathbf{b}_l \circ \mathbf{c}_l, \\ \mathbf{X}^{(m\times np)} &\approx \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T. \end{aligned} \tag{1}$$

The PARAFAC model may be extended to N-way data.

Without loss of generality, we normalize all columns of the factor matrices to unit length and store the accumulated weight in a vector $\boldsymbol{\lambda}$:

$$\mathcal{X} \approx \sum_{l=1}^{r} \lambda_l (\mathbf{a}_l \circ \mathbf{b}_l \circ \mathbf{c}_l).$$

We also impose a constraint on the final solution such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$. Because this normalization and reordering may be performed in a post-processing step, we describe an algorithm for fitting the model without λ.

A common approach to fitting the PARAFAC model to data is an alternating least squares (ALS) algorithm [2, 7, 8], where one cycles over all factor matrices and performs a least-squares update for one factor matrix while holding all the others constant.

Our analysis deals with a variant of the PARAFAC model that constrains the factor matrices to be nonnegative. Often this is called nonnegative tensor factorization (NTF) or sometimes just nonnegative PARAFAC.

The goal of NTF is to find the best fitting nonnegative matrices $\mathbf{A} \in \mathbb{R}_+^{m\times r}$, $\mathbf{B} \in \mathbb{R}_+^{n\times r}$, and $\mathbf{C} \in \mathbb{R}_+^{p\times r}$ in the PARAFAC model, Equation (1), that fit the data in $\mathcal{X}$. This corresponds to the following minimization problem:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \mathcal{X} - \sum_{l=1}^{r} \mathbf{a}_l \circ \mathbf{b}_l \circ \mathbf{c}_l \right\|. \tag{2}$$

To compute the NTF, we solve a series of nonnegative factorization (NMF) subproblems

$$\begin{aligned} &\min_{\mathbf{A}\in\mathbb{R}_+^{m\times r}} ||\mathbf{X}^{(m\times np)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T||_F, \\ &\min_{\mathbf{B}\in\mathbb{R}_+^{n\times r}} ||\mathbf{X}^{(n\times mp)} - \mathbf{B}(\mathbf{C} \odot \mathbf{A})^T||_F, \\ &\min_{\mathbf{C}\in\mathbb{R}_+^{p\times r}} ||\mathbf{X}^{(p\times mn)} - \mathbf{C}(\mathbf{B} \odot \mathbf{A})^T||_F. \end{aligned}$$

Each of these matrix systems is treated as an NMF problem and solved in succession using the multiplicative update introduced by Lee and Seung [9].

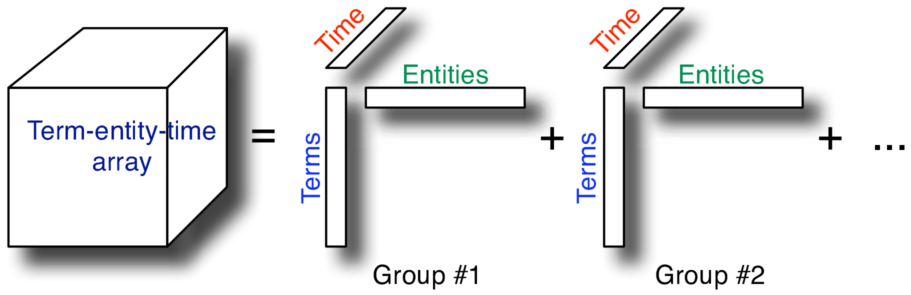


Fig. 1. PARAFAC and the NTF provide a 3-way decomposition for associating terms with entities over time.

This procedure was first used by Welling and Weber [10]. We have adapted their algorithm and incorporated the addition of ϵ for stability as was done in [11, 12]. For example, to solve for $\mathbf{A}$ we compute

$$a_{i\rho} \leftarrow a_{i\rho} \frac{(\mathbf{X}^{(m\times np)}\mathbf{Z})_{i\rho}}{(\mathbf{A}\mathbf{Z}^T\mathbf{Z})_{i\rho}+\epsilon}, \mathbf{Z} = (\mathbf{C}\odot\mathbf{B}).$$

Here, ϵ is a small number like 10^{-9} that adds stability to the calculation and guards against introducing a negative number from numerical underflow. An algorithm for NTF was written in MATLAB, using sparse extensions of the Tensor Toolbox [13, 14].

The approximation rank $r > 0$ of NTF corresponds to the number of associations in the corpus that we are interested in finding. Each triad $\{\mathbf{a}_j, \mathbf{b}_j, \mathbf{c}_j\}$, for $j = 1, \ldots, r$, defines scores for a set of terms, entities, and time for a particular association in the corpus of news stories, as shown in Figure 1. The value λ_j (after normalization) indicates the weight of the association for triad j. Each column of $\mathbf{C}$ records the strength of each association over time.

4 VAST 2007 Dataset

In testing our NTF algorithm from Section 3.2, we processed $1,455$ text files corresponding to news stories, email messages or blog posts from the VAST 2007 Contest dataset [3]. In addition to the plain text versions of these files, the data includes corresponding *tagged* files. The five SGML-based SGML tags are *date*, *person*, *location*, *organization* and *money*. A sample extract from one of the tagged news stories is provided below:

```
Activists performed a skit of the <ENAMEX TYPE="PERSON">Executive
Yuan</ENAMEX> failing to supervise its subordinates to prevent
avian flu.  "The hygiene required when butchering chickens could
not stressed more at this point, where we are exposed to the
potential spread of avian flu," said <ENAMEX TYPE="PERSON">Chen
Yu-min</ENAMEX> (??????), <ENAMEX TYPE="ORGANIZATION"><ENAMEX
TYPE="ORGANIZATION">director of the <ENAMEX TYPE="ORGANIZATION">
Environment and Animal Society</ENAMEX></ENAMEX> of Taiwan
</ENAMEX>.
```

Each file within this data set has a unique time stamp associated with it. The *date* tag was used to help extract the time stamp from each file, but was otherwise ignored. The four remaining tags all represent entities of interest and were used to create a single comma-delimited file that later served as input for the NTF model. Two separate dictionary files[3] were used in the process, one for terms and one for entities. The dictionaries allowed the words to be replaced by numeric IDs. A Python script using the built-in SGML Parser class and the dictionary files described above was used to generate a comma-delimited file of this format:

`Date, Entity Type, Entity ID, Term ID, Term Frequency Count`[4]

A sample extract from the resulting file is given below:

```
Mon Sep  8 08:40:46 2003, MONEY, 786, 2181, 1
Mon Sep  8 08:40:46 2003, MONEY, 786, 6113, 2
Mon Sep  8 08:40:46 2003, ORGANIZATION, 781, 922, 4
Mon Sep  8 08:40:46 2003, ORGANIZATION, 781, 300, 2

Mon Sep  8 08:40:46 2003, LOCATION, 784, 2181, 1
Mon Sep  8 08:40:46 2003, LOCATION, 784, 6113, 2
Mon Sep  8 08:40:46 2003, PERSON, 783, 922, 4
Mon Sep  8 08:40:46 2003, PERSON, 783, 300, 2
```

5 NTF Analysis

In order to answer the questions posed in Section 2.2 for scenario analysis, we considered term-by-entity associations in the news stories over monthly time intervals, which corresponds to a sparse array $\mathcal{X}$ of size $12,121 \times 7141 \times 15$ with

[3] The dictionary files were generated by the General Text Parser (GTP) software environment, see [15].

[4] within a particular file.

1,142,077 nonzeros. We scaled the nonzero term counts according to a weighted frequency:

$$x_{ijk} = l_{ijk} t_i e_j w_k,$$

where l_{ijk} is the local weight for term i co-occurring with entity j in a news story during month k, t_i is the global weight for term i, e_j is the global weight for entity j, and w_j is a time normalization factor.

Let f_{ijk} be the frequency (raw count) of term i appearing in the same news story as entity j during month k. The specific formulas of the scaled tensor are listed here:

$$\begin{array}{ll} \text{Log local weight} & l_{ijk} = \log(1 + f_{ijk}) \\ \text{Term global weight} & t_i = 1 + \sum_{j,k} \frac{h_{ijk} \log h_{ijk}}{\log np}, \quad h_{ijk} = \frac{f_{ijk}}{\sum_{jk} f_{ijk}} \\ \text{Entity global weight} & e_j = 1 + \sum_{i,k} \frac{g_{ijk} \log g_{ijk}}{\log mp}, \quad g_{ijk} = \frac{f_{ijk}}{\sum_{ik} f_{ijk}} \\ \text{Time normalization} & w_k = \frac{1}{\sqrt{\sum_{i,k} (l_{ijk} t_i e_j)^2}} \end{array}$$

The global weights are adapted from the well-known log-entropy weighting scheme [16] used on term-by-document matrices. The log local weight scales the raw term frequencies to diminish the importance of high frequency terms. The entropy global weight helps to discriminate important terms and entities from frequently occuring terms and entities. The time normalization helps to correct imbalances in the number news stories over each time period.

For the VAST 2007 Contest data (see Section 2), we computed a 25-component ($r = 25$) nonnegative decomposition of the term-entity-month array $\mathfrak{X}$. One iteration took about 25 seconds on a 3GHz Pentium Xeon desktop computer with 2GB of RAM. Most runs required about 20 iterations to satisfy a tolerance of 10^{-4} in the relative change of fit. We chose the best minimizer from among ten runs starting from random initializations. The relative norm of the difference for the one chosen was 0.8881.

5.1 Model Outputs and Postprocessing

Processing the data with NTF resulted in twenty-five groups consisting of interrelated entities and terms. Fifteen entities of various types and thirty-five terms describe each group. The following excerpt from the NTF output demonstrates the layout of Group 20 (Environmental and animal rights issues in China), see Table 1:

```
############ Group 20  ##########
Scores        Idx  Name
 0.2252609   4680 scott roberts
 0.2252609   4685 zhang
 0.2252609   4687 roberts
 0.2252609   4682 iron and steel statistics bureau
 0.2252609   4686 $13 billion
 ...
 0.2252609   4683 zhang jianyu
 0.2252609   4690 kazakhstan
 0.2048428    799 massachusetts
 0.1827936    259 brazil
 0.1740706     77 beijing

Scores        Idx  Term
 0.2140977   3644 energy
 0.1915396   1855 china
 0.1502502   8104 power
 0.1321501   1011 beijing
 0.1239235   7340 oil
 0.1155490   9140 roberts
 0.1130895   2146 communist
 0.1129717  10203 steel
 0.1057306   2023 coal
```

For each of the twenty-five groups, a relevance score was computed for every story in the data set. Entities were given twice the weight of terms for the purpose of this computation. In order to improve the chances of smaller-length documents being judged relevant, *Lnu scaling* [16] was applied. If $A = [a_{ij}]$ is the term-by-document where by each element a_{ij} defines the relevancy of term (or entity) i to document (or story) j, then we define

$$a_{ij} = \frac{(1 + \log(f_{ij})) \,/\, \left(1 + \log(\bar{f}_{ij})\right)}{(1 - s) \times p + s \times u}, \text{ where } \bar{f}_{ij} = \frac{\sum\limits_i f_{ij}}{\sum\limits_i \chi(f_{ij})}.$$

The slope s is set to 0.2; p,the average number of unique words in a document within the data set was determined to be 158; and u, the number of unique words in document or story j was calculated a priori. The function χ is defined as

$$\chi(r) = \begin{cases} 1 \text{ if } r > 0, \\ 0 \text{ if } r = 0, \end{cases}$$

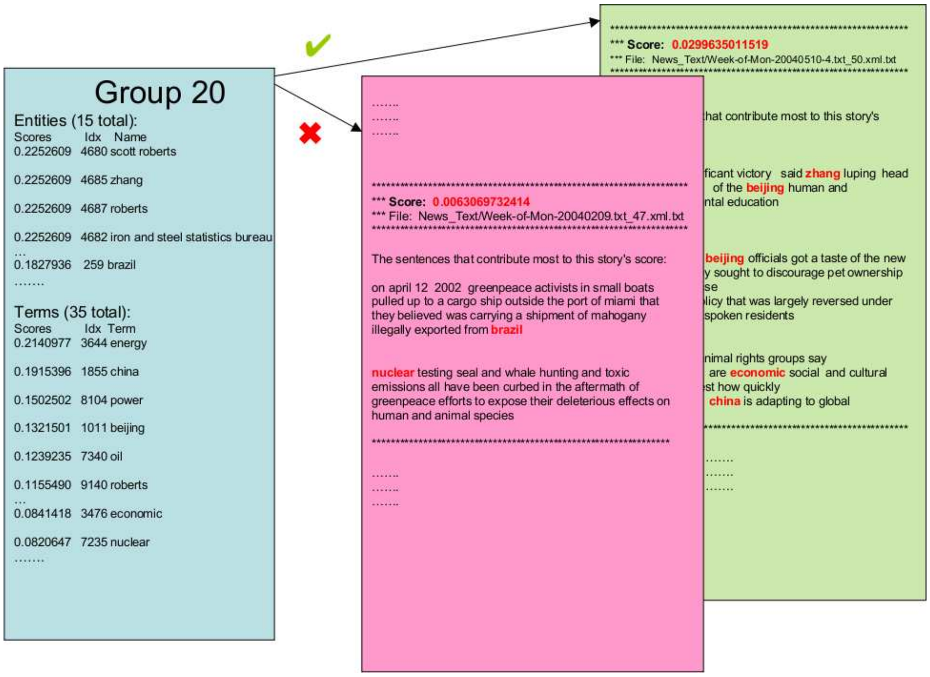
so that the overall relevance score R for each document (story) j using the terms and entities (indexed by i) of the given tensor group is simply defined by

$$R = \sum_i a_{ij}. \tag{3}$$

For relevant stories, in addition to the overall score, a score was also computed for every individual sentence. For each story, the three top-scoring sentences were then printed out as a brief summary. This significantly improved the analysis process by decreasing the amount of time necessary to determine a story's subject and whether it was in fact relevant to a given tensor group. In addition to the top three sentences, each summary also contains the overall story score and a reference to the file containing the full story. The latter was used to extract events and activities from the story. The events and activities were listed in a spreadsheet in chronological order. The scoring and summarization of two different news stories against the Group 20 (see Table 1) is illustrated in Figure 2.

An empirically determined threshold (0.0088) on R was used to create a set of relevant stories. In some cases, no further processing was necessary. In other cases, where many clearly irrelevant documents were included in the set, a slightly higher secondary threshold value was applied. Table 2 shows the reduction in news stories associated with a given NTF-generated Group as the threshold on R is increased.

The extraction of events and activities was a somewhat subjective process, involving occasional ambiguity between events and activities. In general, a discrete, one-time action was typically classified as an event, while a continuous action was classified as an activity.



```
Group 20
Entities (15 total):
Scores      Idx   Name
0.2252609   4680 scott roberts

0.2252609   4685 zhang

0.2252609   4687 roberts

0.2252609   4682 iron and steel statistics bureau
...
0.1827936    259 brazil
.......

Terms (35 total):
Scores      Idx  Term
0.2140977   3644 energy

0.1915396   1855 china

0.1502502   8104 power

0.1321501   1011 beijing

0.1239235   7340 oil

0.1155490   9140 roberts
...
0.0841418   3476 economic

0.0820647   7235 nuclear
.......
```

```
.......
.......
.......


******************************************************************
*** Score:  0.0063069732414
*** File:  News_Text/Week-of-Mon-20040209.txt_47.xml.txt
******************************************************************

The sentences that contribute most to this story's score:

on april 12  2002  greenpeace activists in small boats
pulled up to a cargo ship outside the port of miami that
they believed was carrying a shipment of mahogany
illegally exported from brazil


nuclear testing seal and whale hunting and toxic
emissions all have been curbed in the aftermath of
greenpeace efforts to expose their deleterious effects on
human and animal species

****************************************************************

.......
.......
.......
```

```
******************************************************************
*** Score:  0.0299635011519
*** File:  News_Text/Week-of-Mon-20040510-4.txt_50.xml.txt
******************************************************************

hat contribute most to this story's


ficant victory   said zhang luping  head
 of the beijing human and
ntal education


beijing officials got a taste of the new
y sought to discourage pet ownership
se
licy that was largely reversed under
spoken residents


nimal rights groups say
 are economic social  and cultural
st how quickly
 china is adapting to global

*****************************************

.......
.......
.......
```

Fig. 2. Demonstration of scoring news stories against terms and entities from Tensor Group 20 (Environmental and animal rights issues in China).

Table 1. Themes of identifiable activities and events from the NTF model.

Group	Theme (topics)
1	Conservation of large feliness (tigers and leopards)
4	Chinese government's actions related to environmental/animal rights
6	Disease: potential dangers of genetic modification related to food-borne illnesses; animal-borne diseases and their transfer to humans; effects of diet on health
9	Exotic animal and drug trafficking
10	Meat alternatives and their benefits for consumer health; animal treatment standards
11	Spread of Animal diseases to humans
13	Animal-rights groups responsible for attacks on pet stores and supermarkets; People for the Ethical Treatment of Animals (PETA)
15	Monkeypox outbreak in the United States
16	Research on chimpanzees and other primates in the wild
18	Activities by the Earth Liberation Front (ELF) and similar organizations
20	Environmental and animal rights issues in China
23	Animal fighting for entertainment (e.g., bullfighting and cockfighting)

Table 2. Threshold effects on event/activity extraction.

	Threshold (Article Count)	
Group	Primary	Secondary
1	0.0088 (59)	0.0147 (24)
4	0.0088 (156)	0.0190 (21)
6	0.0088 (42)	
9	0.0088 (1019)	0.0410 (10)[a]
10	0.0088 (219)	0.0210 (27)
11	0.0088 (401)	0.0210 (14)
13	0.0088 (700)	0.0310 (16)
15	0.0088 (1283)	0.0350 (19)
16		
18	0.0088 (65)	0.0147 (35)
20	0.0088 (9)	
23	0.0088 (22)	0.0210 (5)

[a] Additional weighting for the terms *fish*, *cocaine*, and *trafficking* was needed to obtain this smaller subset of related stories.

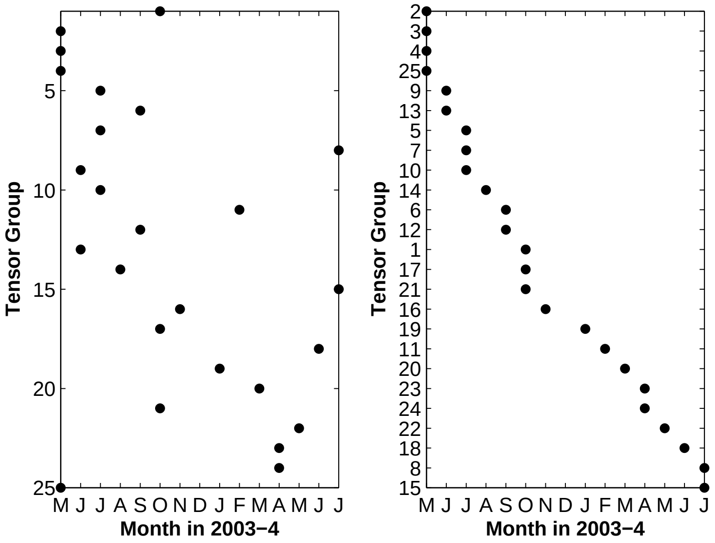


Fig. 3. Peak month by tensor group matrix from the NTF model; original ordering (left) and re-ordered matrix (right).

5.2 Scenario Discovery

Recalling the *nature* of the scenario defining the VAST 2007 Contest dataset (see Section 2.1), we note that Groups 6, 9, 15, and 18 from Table 1 expose both events and activities of two important terrorist plots: cocaine trafficking via tropical fish trade and spread of monkeypox by infected chinchillas. Examining the peak nonnegative components of each column of the **C** factor of our NTF model (see Section 3.2) we can sequence the Groups of Table 1 through time and show (see Figure 3) the progression of the events and activities through the timespan of the entire dataset. In Figure 4 we highlight the **C**-factor *spikes* of the groups associated with the dominant terrorist plots of the VAST 2007 Contest [3]. Tables **??** and 4 in the Appendix list the dominant terms and entities of Groups 9 and 15 from Table 1 along with the top-scoring events or activities for each group. The descriptions of each event (or activity) were hand-curated from the top scoring sentences (see Section 5.1) of each relevant story.

6 Conclusions

We have demonstrated how a nonnegative tensor factorization (NTF) model can be used to generate useful term-by-entity associations from textual-based data

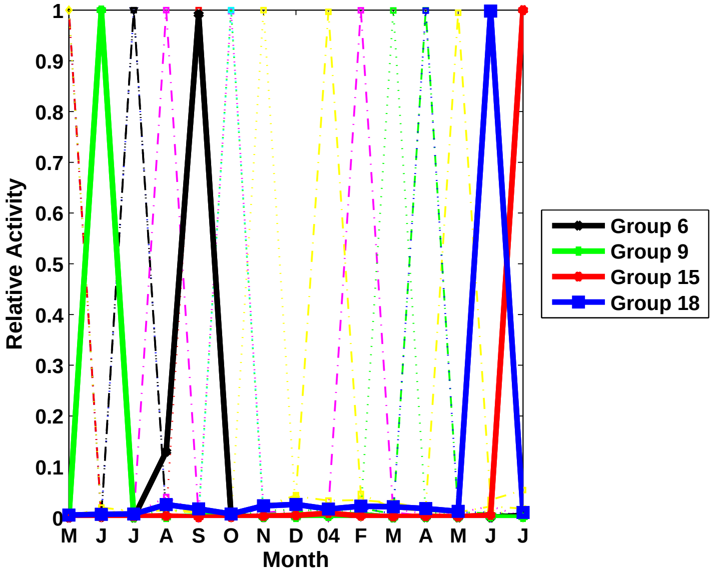


Fig. 4. Plot of all 25 time-based tensor groups (**C**-factor) generated by the NTF model; groups 6,9,15,18 (identifying ecoterrorism events and activities) are highlighted.

(news stories, blog entries, etc.). Using the VAST 2007 Contest dataset and its fictitious terrorist plots, we have generated interpretable tensor factors that can be used to define and track events and activities critical to those plots. These preliminary results are quite promising and suggest that NTF may become an effective tool for the emerging discipline of visual analytics. Further research in the effects of scaling/weighting of the attributes defining the inital tensor (or datacube) is needed along with the incorporation of the underlying NTF model into a visual interface more suitable for human steering and querying.

References

1. Bader, B.W., Berry, M.W., Browne, M.: Discussion tracking in Enron email using PARAFAC. In: Survey of Text Mining II: Clustering, Classification and Retrieval. Springer-Verlag, London (2008) 147–163
2. Harshman, R.A.: Foundations of the PARAFAC procedure: models and conditions for an “explanatory” multi-modal factor analysis. UCLA working papers in phonetics **16** (1970) 1–84 Available at `http://publish.uwo.ca/~harshman/wpppfac0.pdf`.
3. Scholtz, J., Plaisant, C., Grinstein, G.: IEEE VAST 2007 Constest. `http://www.cs.umd.edu/hcil/VASTcontest07` (2007)
4. Smilde, A., Bro, R., Geladi, P.: Multi-way analysis: applications in the chemical sciences. Wiley, West Sussex, England (2004)

5. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. Psychometrika **35** (1970) 283–319
6. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review To appear.
7. Faber, N.K.M., Bro, R., Hopke, P.K.: Recent developments in CANDECOMP/PARAFAC algorithms: A critical review. Chemometrics and Intelligent Laboratory Systems **65**(1) (January 2003) 119–137
8. Tomasi, G., Bro, R.: A comparison of algorithms for fitting the PARAFAC model. Computational Statistics & Data Analysis **50**(7) (April 2006) 1700–1734
9. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (21 October 1999) 788–791
10. Welling, M., Weber, M.: Positive tensor factorization. Pattern Recognition Letters **22**(12) (2001) 1255–1261
11. Berry, M.W., Browne, M.: Email surveillance using nonnegative matrix factorization. In: Workshop on Link Analysis, Counterterrorism and Security, SIAM Conference on Data Mining, Newport Beach, CA (2005)
12. Berry, M., Browne, M.: Email Surveillance Using Nonnegative Matrix Factorization. Computational & Mathematical Organization Theory **11** (2005) 249–264
13. Bader, B.W., Kolda, T.G.: Efficient MATLAB computations with sparse and factored tensors. Technical Report SAND2006-7592, Sandia National Laboratories, Albuquerque, New Mexico and Livermore, California (December 2006)
14. Bader, B.W., Kolda, T.G.: Matlab tensor toolbox, version 2.1. `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/` (December 2006)
15. Giles, J., Wo, L., Berry, M.: GTP (General Text Parser) Software for Text Mining. In Bozdogan, H., ed.: Software for Text Mining, in Statistical Data Mining and Knowledge Discovery. CRC Press, Boca Raton, FL (2003) 455–471
16. Berry, M., Browne, M.: Understanding Search Engines: Mathematical Modeling and Text Retrieval. Second edn. SIAM, Philadelphia, PA (2005)

Appendix

The dominant terms and entities (based on nonnegative components of factors **A** and **B** from Equation (1)) for Groups 9 and 15 (see Table 1) along with the top-scoring news stories (with summaries) using thresholds on the R measure defined by Equation (3).

Table 3. Group 9 (tropical fish and drug trafficking) highest-ranked entities and terms along with the events of the high-scoring news stories for that group; the *Evidence* column contains quotes from the original article text and the *Dateline* format is mm/dd/yy.

Intensity	**Entity**	**Entity**	**Intensity**	
Score	**Name**	**Type**	**Score**	**Term**
0.3588235	$215 million	monetary	0.3258677	banks
0.3588235	cruz	location	0.2219373	fishes
0.3588235	darla banks	person	0.1687334	tropical
0.3588235	$25-30 million	monetary	0.1465103	trafficking
0.3218558	rio de janeiro	location	0.1447117	brazil
0.2687660	alabama	location	0.1373243	illegal
0.2654457	brazil	location	0.1254398	poachers
0.2524907	south america	location	0.1246595	fish
0.1765029	us	location	0.1229337	trade
0.0300772	north america s	location	0.1147486	sinister
0.0300763	$22 billion	monetary	0.1145697	frontline
0.0300747	cspi s dewaal	person	0.1103388	janeiro
0.0300746	jean chretien	person	0.1103138	concealed
0.0300697	hansen	person	0.1103130	trophies

Event	**R-Score**	**Dateline**	**Description**
1	0.04138	10/27/03	Shipping bags for tropical fish are tainted
2	0.04138	10/27/03	Company (Global Ways) blames packer
3	0.05753	01/06/04	Warning on fish imports from South America

Event	**Evidence**
1	"I urge your readers to not use Global Ways...low survival rates, poor packaging, poor responsiveness. We will never use them again... the packaging on my last shipment of Tigrinus Catfish was abominable. Shipping bags were covered in some noxious substance that caused our handlers hands to go numb and eventually need emergency medical treatment. You can imagine the condition of the fish themselves. Less than 10% survived their trip."
2	"In response to these comments, Global Ways spokeman blamed an inexperienced packer in South America for problems in a very few shipments. We are working with these customers and have guaranteed them satisfaction. A problem occurred in a very small number of the 5000 shipments we handled last quarter."
3	"The Fish and Wildlife Service (FWS) is warning ornamental fish merchants not to handle any shipments of tropical catfish from South America that may have come into the United States through Miami, as the packaging bags may be contaminated with a substance that can cause serious illness. The source of the toxin is unknown, but it is not directly related to the imported fish themselves."

Table 4. Group 15 (monkeypox outbreak) highest-ranked entities and terms along with the events of the high-scoring news stories for that group; the *Evidence* column contains quotes from the original article text and the *Dateline* format is mm/dd/yy.

Intensity	**Entity**	**Entity**	**Intensity**	
Score	**Name**	**Type**	**Score**	**Term**
0.2485621	bruce longhorn	person	0.2958673	monkeypox
0.2485621	david chelmsworth	person	0.2054770	outbreak
0.2485621	cesar gil	person	0.2008147	longhorn
0.2485621	virginia tech	organization	0.1594331	gil
0.2485621	mary ann ollesen	person	0.1552401	chinchilla
0.2485621	mouse club of america	organization	0.1434742	travel
0.2485621	westminster	location	0.1391984	sars
0.2485621	la	location	0.1379675	chinchillas
0.2240710	rat	organization	0.1342139	continent
0.2138531	college of veterinary		0.1294389	expect
	medicine	organization	0.1215461	sick
0.1960454	centers for disease control		0.1161760	outbreaks
	and prevention	organization	0.1144558	exotic
0.1878988	atlanta	location	0.1122925	pets
			0.1026513	pot-bellied

Event	***R* Score**	**Dateline**	**Description**
1	0.04119	05/15/04	Monkeypox discovered in the Wisconsin, Illinois, Indiana
2	0.04119	07/07/04	Monkeypox cases found in the Los Angeles area
3	0.14698	07/24/04	Two people die of monkeypox. Chinchillas related to spread of virus
4	0.14698	07/24/04	Cesar Gil wanted by authorities in connection with the monkeypox outbreak

Event	**Evidence**
1	“The disease, which is not usually fatal to humans, has previously been discovered in Wisconsin, Illinois and Indiana in mid-May. Most of those affected had contact with prairie dogs - a type of wild rodent which lives in burrows on the western US plains, and sold as pets.”
2	“Seven people in the Los Angeles area have seriously ill from monkeypox - a rare smallpox-like disease, US health officials say. More than 50 possible cases are being investigated.”
3	“But as the monkeypox outbreak shows, there are risks in owning exotic pets. Now that two people have died from a new, more potent strain, the mishmash of regulations leaves loopholes that have allowed sick animals to travel, health officials say. Chinchillas, believed to be the cause of the recent outbreak and on the endangered species lists, are part of that concern.”
4	“Cesar Gil, an animal rights activist and chinchilla breeder in the LA area, is currently being sought in connection with the monkeypox outbreak. He has also been a suspect in radical animal rights activities. Officials believe Gil has fled the country.”