A Visual Approach to Automated Text Mining and Knowledge Discovery

Doctoral Dissertation
by
Andrey A. Puretskiy
Advisor: Dr. Michael W. Berry

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

November 5, 2010

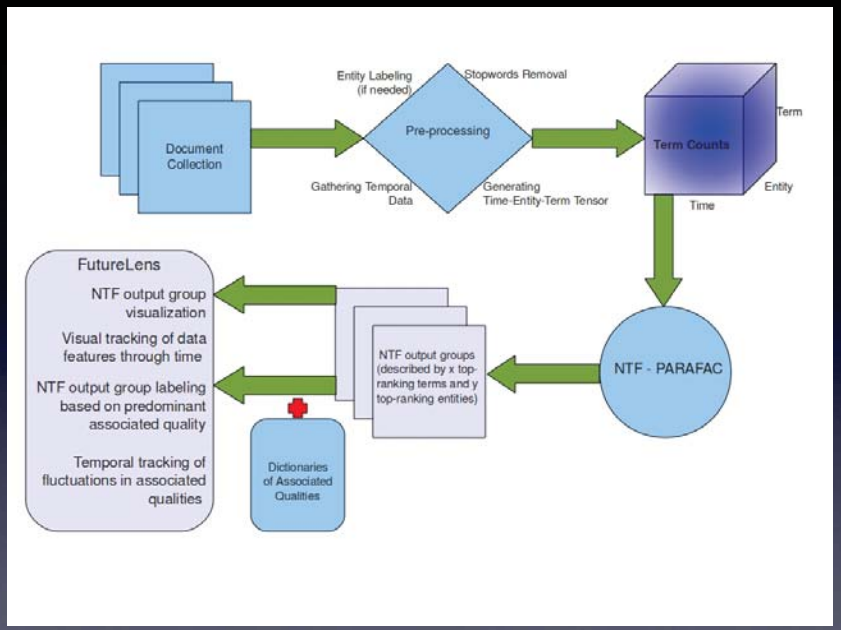
Motivations

- Vast Quantities of Text Available
  - Scientific Literature
  - News Articles and Blogs
  - Email
- Effective Visual Analytics Requirements:
  - Process Vast Quantities of Textual Information
  - Significant Automation of Analysis
  - Visual, Human-understandable Results Presentation

Dissertation Proposal Revisited

- Integrate visual post-processing and nonnegative tensor factorization (NTF)
- Improve upon existing NTF technique
  - Allow the user to affect factorization by adjusting term weights within the tensor
  - Add automated result classification to visual results post processing
- Demonstrate effectiveness of approach using several different datasets
- Create an environment for testing of different heuristics for tensor rank estimation

Visual Analytics Environment Architecture



Tensor Factorization

- Tensor: Multidimensional array
- History: Hitchcock (1927), Cattell (1944), Tucker (1966)
- Factorization: Process of rewriting a tensor as a finite sum of lower-rank tensors
- PARAFAC: Parallel Factors Analysis (Harshman, 1970)

Tensor Factorization: PARAFAC Methodology

- Given tensor $\mathcal{X}$ and rank R, define the factor matrices as combinations of vectors from rank-one components

$$\mathcal{X} \approx A \circ B \circ C \approx \sum_{r=1}^{R} a_r \circ b_r \circ c_r$$

- Alternating Least Squares:

  Cycle "over all the factor matrices and perform a least-squares update for one factor matrix while holding all the others constant." (Bader, 2008)

Tensor Factorization - Summary

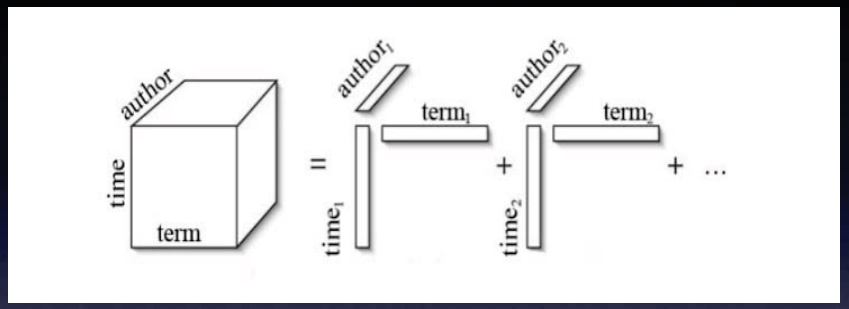


Illustration of a Time-by-Author-by-Term Tensor Decomposition

Nonnegative Tensor Factorization (NTF)

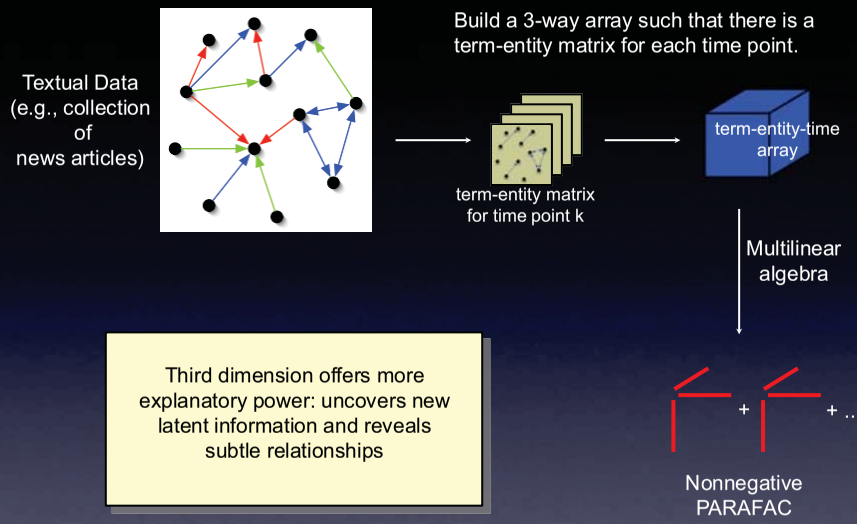
- Nonnegative tensor factorization algorithm: PARAFAC with nonnegativity constraint
- Matlab® Code (Dr. Brett Bader, Sandia)
- Python Translation (Mr. Papa Diaw, Advisor: Dr. Michael Berry)
- Extracts features from textual data
- Each feature may be described by a list of terms and tagged entities

Performance Comparison

Dataset	Number of Documents	Avg. Document Length (terms)	Matlab NTF Execution Time (minutes)	Python NTF Execution Time (minutes)
Kenya 2001-2009	900	696	4.54	17.15
VAST 2007	1455	391	3.95	16.13

- Times were averaged over 10 trials
- While not as fast as Matlab®, Python still allows real-time analysis
- Future improvements in Python NTF code performance may be possible

NTF: Multidimensional Data Analysis



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

Sample NTF Output

```
############ Group 15 ##########
Scores      Idx  Name
0.2485621   7120 bruce longhorn  7120
0.2485621   7122 longhorn  7122
0.2485621   7128 chelmsworth  7128
0.2485621   7124 gil  7124
0.2485621   7121 virginia tech  7121
0.2485621   7125 mary ann ollesen  7125
…
Scores      Idx  Term
0.2958673   6907 monkeypox
0.2054770   7468 outbreak
0.2008147   6358 longhorn
0.1594331   4644 gil
0.1552401   1856 chinchilla
0.1434742  11049 travel
0.1391984   9322 sars
0.1379675   1857 chinchillas
0.1342139   2372 continent
0.1294389   3888 expect
0.1215461   9711 sick
0.1161760   7469 outbreaks
0.1144558   3883 exotic
0.1122925   7824 pets
0.1026513   8088 pot-bellied
0.1026513   7229 novelty
0.1019125   1742 cesar
0.1004109  10280 strain
0.1000808   5878 jul
…
```


FutureLens
Features

- Automatically Loads All Terms Found in Input Dataset (except those on the list of exclusions)
- Ability to Search through Terms
- Ability to Sort Terms
- Ability to Create Collections of Terms
- Ability to Create Phrases
- A more complete description of capabilities and effectiveness published in: G.L. Shutt, A.A. Puretskiy, M.W. Berry: *FutureLens: Software for Text Visualization and Tracking*. Text Mining Workshop, Proceedings of the Ninth SIAM International Conference on Data Mining, Sparks, NV, April 30-May 2, 2009, ISBN: 978-0-898716-82-5.

Completed Goals

- Integration of Pre-processing, NTF, and FutureLens into a single analysis environment
- Allowing the user to affect the NTF process through Integrated Analysis Environment controls:
  - User is able to define relative importance (or trustworthiness) of terms or subsets of terms
- Introduction of automatic NTF results classification through the use of pre-existing and user-modifiable dictionaries

Integrated Analysis Environment

Features and Design Objectives

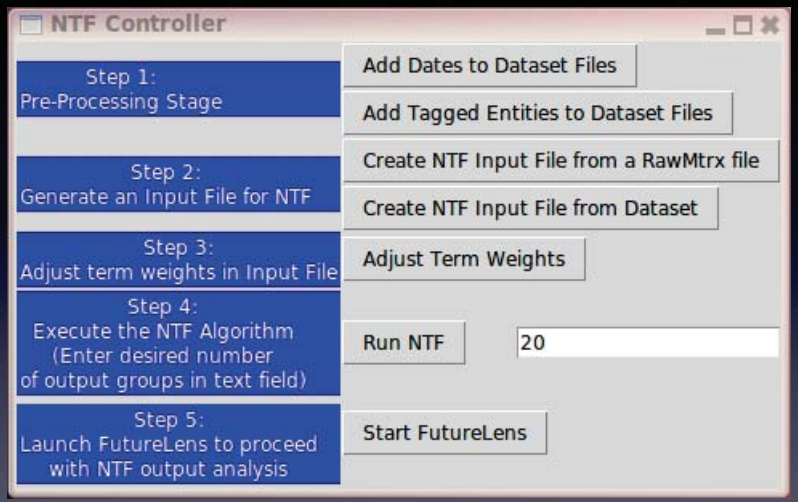
- Objectives
  - A single application
  - Simple look to avoid feature overload
  - Easy to use without much experience
  - Integration of multiple important capabilities
- Implemented in Python
  - Portability
    - Linux, OS X, Windows
  - Look and feel of application native to the user's operating system
  - Easily modifiable due to Python's excellent readability

Integrated Analysis Environment

Capabilities

- Addition of temporal information into the dataset in SGML-tagged format
- User-customized entity tagging (SGML format)
- NTF input file creation
- Tensor term weight adjustment
- Python NTF PARAFAC execution
- FutureLens launching for continuing visual analysis of NTF results

Integrated Analysis Environment



Tensor Term Weights Adjustment
Motivation

- Lack of interest in subset of terms
  - Terms may have been deemed "untrustworthy"
  - Terms may likely be irrelevant to particular analysis model
  - The above may be insufficient to eliminate terms as stopwords
- Strong interest in a subset of terms
  - Subset may have been deemed particularly trustworthy
  - Analyst may need to create a model that focuses strongly on a particular aspect of the data

Tensor Term Weights Adjustment
The Simple Approach

- Plain-text files containing lists of terms
  - Easy for computer-inexperienced users
  - Each file corresponds to a particular analysis model
  - Very easy to create, distribute, view, share feedback, modify models
  - Integrated Analysis Environment quickly creates a term-weight modified NTF input file based on such input

Automated NTF Output Group Labeling

- Motivation: Increase efficiency of human analysis of NTF results

- Automated labeling feature functions much faster than analyst labeling ever could

- Feature allows the analyst to quickly sort NTF output groups by analyst-defined categories
  - Focus exclusively on category or categories of interest
  - Feature includes a default ("none of the above") category

Automated Labeling
Design and Utilization

- Plain-text files containing lists of terms
  - Easy for computer-inexperienced users
    - Very easy to create, distribute, view, share feedback, modify models

- FutureLens quickly labels NTF output groups based on the set of category descriptor files loaded at the time

- Visual category labeling allows the analyst to filter out uninteresting groups and focus on the ones most pertinent to the focus of analysis

Integrated Analysis Environment Demo

Conclusions

- The demonstrated approach can be effectively used to analyze vast quantities of textual data
- The approach is straightforward and easy to use even for computer-inexperienced analysts
- The approach is highly portable and functions under Linux, OS X, and Windows

Future Research Directions

- Integration of Spatial Information
  - Geo-coding
  - Allow the user to track term usage changes and fluctuations through geographical locales
- Bioinformatics applicability
  - Medical research literature
  - Gene-by-Term-by-Expression data may reveal additional functional relationships among genes

References

- Brett W. Bader, Andrey A. Puretskiy, and Michael W. Berry. *Scenario Discovery Using Nonnegative Tensor Factorization*. In Jose Ruiz-Shulcloper and Walter G. Kropatsch, editors, Progress in Pattern Recognition, Image Analysis and Applications, Proceedings of the Thirteenth Iberoamerican Congress on Pattern Recognition, CIARP 2008, Havana, Cuba, Lecture Notes in Computer Science (LNCS) 5197, pages 791–805. Springer-Verlag, Berlin, 2008.
- G.L. Shutt, A.A. Puretskiy, M.W. Berry: *FutureLens: Software for Text Visualization and Tracking*. Text Mining Workshop, Proceedings of the Ninth SIAM International Conference on Data Mining, Sparks, NV, April 30-May 2, 2009, ISBN: 978-0-898716-82-5.
- A.A. Puretskiy, G.L. Shutt, and M.W. Berry, "Survey of Text Visualization Techniques," in Text Mining:
- Applications and Theory, M.W. Berry and J. Kogan (Eds.), Wiley, Chichester, UK, pp. 107-127, 2010.

References

- Exploring and visualizing frequent patterns in text collections with FeatureLens. http://www.cs.umd.edu/hcil/textvis/featurelens. Visited November 2008.
- Brett W. Bader, Michael W. Berry, and Murray Brown. Discussion tracking in Enron email using PARAFAC. In M.W. Berry and M. Castellanos, editors, Survey of Text Mining II: Clustering, Classification, and Retrieval, pages 147–163. Springer-Verlag, London, 2008.
- T. Kolda, B. Bader, Tensor Decompositions and Applications. SIAM Review, Vol. 51, No. 3, 2009, pp. 455-500.
- Zhe, X. & Boucouvalas, A.C., 2002. Text-to-Emotion Engine for Real Time Internet Communication. International Symposium on Communication Systems, Networks and DSPs, 15-17 July 2002, Staffordshire University, UK, pp 164-168.
- SEASR. Sentiment Tracking from UIMA Data. http://seasr.org/documentation/uima-and- seasr/sentiment-tracking-from-uima-data/. Visited May 2010.

References

- Brett W. Bader, Michael W. Berry, and Amy N. Langville. Nonnegative matrix and tensor factorization for discussion tracking. In A. Srivastava and M. Sahami, editors, Text Mining: Theory, Applications, and Visualization. Chapman & Hall / CRC Press, 2008.
- A. Don, E. Zhelev, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: integrating text mining with visualization. HCIL Technical report 2007-08, May 2007.
- Tjioe E, Berry MW, Homayouni R. Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization). BMC Bioinformatics. 2010 Oct 7;11 Suppl 6:S14. PMID: 20946597
- Harshman, R.A.: Foundations of the PARAFAC procedure: models and conditions for an “explanatory” multi-modal factor analysis. UCLA working papers in phonetics 16, 1-84 (1970), http://publish.uwo.ca/~harshman/wpppfac0.pdf

Questions?