

Interactive Text Mining with Iterative Denoising

Kendall Giles, PhD

kegiles@vcu.edu

www.people.vcu.edu/~kegiles

Assistant Professor

Department of Statistics and Operations Research

Virginia Commonwealth University

Interactive Text Mining with Iterative Denoising

Example 1: Find SPAM

Subject: driesbound wrote: Did you ...
From: Palisha Philibert
Date: 2/4/07 4:35PM
To: kendall@orionsarrow.com

Now HPGI is preparing for the gold extraction. A lot of specialists are working on the field and are making preparations for operative and active working.

"We are delighted with the acquisition of Orion and we feel the property has excellent potential to produce results that will exceed our expectations," commented Ted Pomerleau, President and Chairman of Hemisphere Gold.

As you know the State Department also noted that Suriname's efforts in recent years to liberalize economic policy created new possibilities for U.S. exports and investments. More over in the situation of changeable global economy more countries start to buy gold for their reserves. But what is important is that a start has been made in buying in the market. New information from HPGI will be at an early date. Also company are starting to hire staff for mine gold region. Underplay is our style (HPGI)

a2u903oyselic1k2wfozw3delm08el8dim8mid
70736671746E7366777C68726A69714577747877743374
MEIWP4PEISDEU0XWRH8PA4B5QCXIGD6A0JBZVP8TK7DUUCZ

Subject: e-Society 2007: CFP (submissions: 26 February 2007)
From: Carla Sa
Date: 2/7/2007 3:45PM
To: kendall@orionsarrow.com

Apologies for cross-postings. Please send to interested colleagues and students

-- CALL FOR PAPERS - Deadline for submissions: 26 February 2007 --

IADIS INTERNATIONAL CONFERENCE E-SOCIETY 2007 Lisbon, Portugal, 3 to 6 July 2007 (<http://www.esociety-conf.org/>) part of the IADIS Multi Conference on Computer Science and Information Systems (MCCSIS 2007) Lisbon, Portugal, 3 to 8 July 2007 (<http://www.mccsis.org>)

* Keynote Speaker (confirmed) Professor Mireia Fernández Ardvol, Universitat Oberta de Catalunya (UOC), Barcelona, Spain

* Conference Background and Goals The IADIS e-Society 2007 conference aims to address the main issues of concern within the Information Society. This conference covers both the technical as well as the non-technical aspects of the Information Society. Broad areas of interest are eGovernment / eGovernance, eBusiness / eCommerce, eLearning, eHealth, Information Systems, and Information Management. These broad areas are divided into more detailed areas (see below). However innovative contributions that don't fit into these areas will also be considered since they might be of benefit to conference attendees.

* Format of the Conference The conference will comprise of invited talks and oral presentations. The proceedings of the conference will be published in the form of a book and CD-ROM with ISBN, and will be available also in the IADIS Digital Library (accessible on-line). The best paper authors will be invited to publish extended versions of their papers in the IADIS Journal on WWW/Internet (ISSN: 1645-7641) and other selected Journals.

Example 1: Find SPAM

Subject: driesbound wrote: Did you ...
From: Palisha Philibert
Date: 2/4/07 4:35PM
To: kendall@orionsarrow.com

Now HPGI is preparing for the gold extraction. A lot of specialists are working on the field and are making preparations for operative and active working.

"We are delighted with the acquisition of Orion and we feel the property has excellent potential to produce results that will exceed our expectations," commented Ted Pomerleau, President and Chairman of Hemisphere Gold.

As you know the State Department also noted that Suriname's efforts in recent years to liberalize economic policy created new possibilities for U.S. exports and investments. More over in the situation of changeable global economy more countries start to buy gold for their reserves. But what is important is that a start has been made in buying in the market. New information from HPGI will be at an early date. Also company are starting to hire staff for mine gold region. Underplay is our style (HPGI)

a2u903oyselic1k2wfozw3delm08el8dim8mid
70736671746E7366777C68726A69714577747877743374
MEIWP4PEISDEU0XWRH8PA4B5QCXIGD6A0JBZVP8TK7DUUCZ

Subject: e-Society 2007: CFP (submissions: 26 February 2007)
From: Carla Sa
Date: 2/7/2007 3:45PM
To: kendall@orionsarrow.com

Apologies for cross-postings. Please send to interested colleagues and students

-- CALL FOR PAPERS - Deadline for submissions: 26 February 2007 --

IADIS INTERNATIONAL CONFERENCE E-SOCIETY 2007 Lisbon, Portugal, 3 to 6 July 2007 (<http://www.esociety-conf.org/>) part of the IADIS Multi Conference on Computer Science and Information Systems (MCCSIS 2007) Lisbon, Portugal, 3 to 8 July 2007 (<http://www.mccsis.org>)

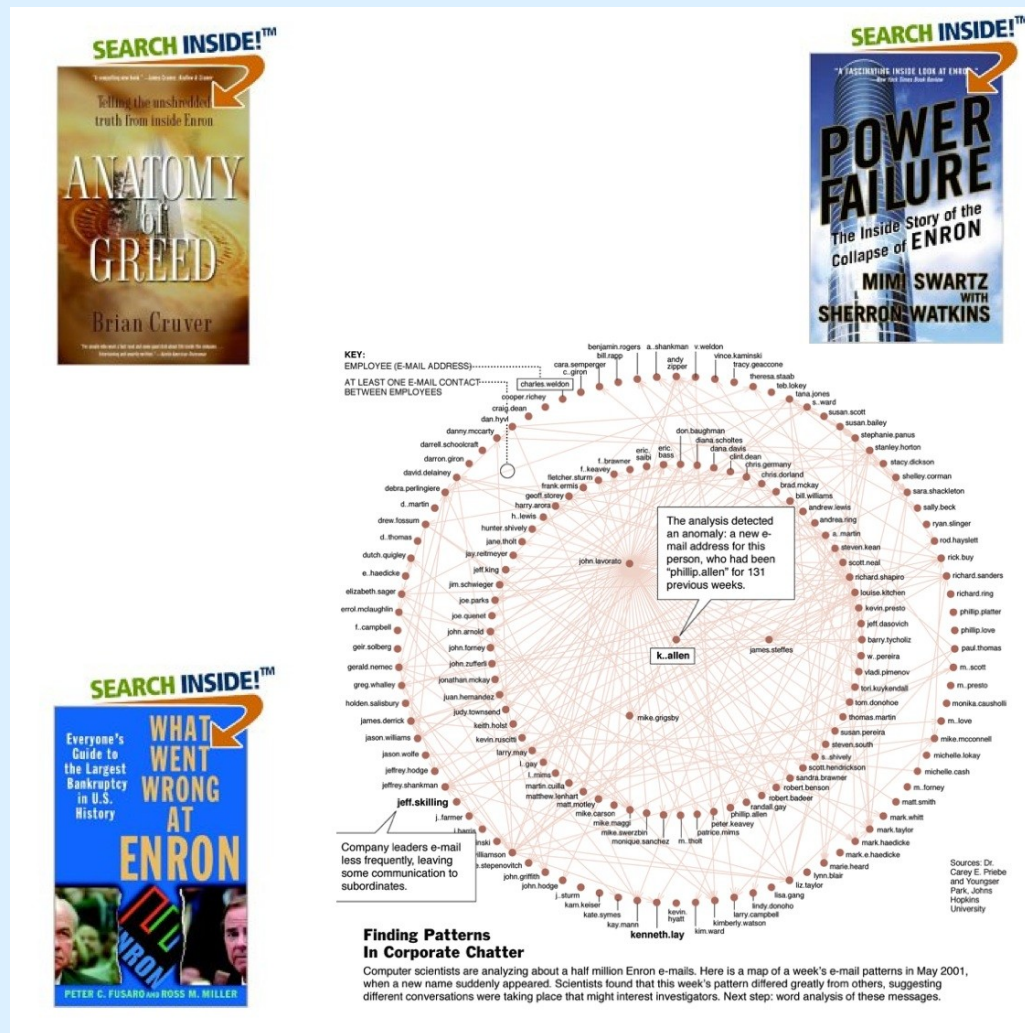
* Keynote Speaker (confirmed) Professor Mireia Fernández Ardvol, Universitat Oberta de Catalunya (UOC), Barcelona, Spain

* Conference Background and Goals The IADIS e-Society 2007 conference aims to address the main issues of concern within the Information Society. This conference covers both the technical as well as the non-technical aspects of the Information Society. Broad areas of interest are eGovernment / eGovernance, eBusiness / eCommerce, eLearning, eHealth, Information Systems, and Information Management. These broad areas are divided into more detailed areas (see below). However innovative contributions that don't fit into these areas will also be considered since they might be of benefit to conference attendees.

* Format of the Conference The conference will comprise of invited talks and oral presentations. The proceedings of the conference will be published in the form of a book and CD-ROM with ISBN, and will be available also in the IADIS Digital Library (accessible on-line). The best paper authors will be invited to publish extended versions of their papers in the IADIS Journal on WWW/Internet (ISSN: 1645-7641) and other selected Journals.

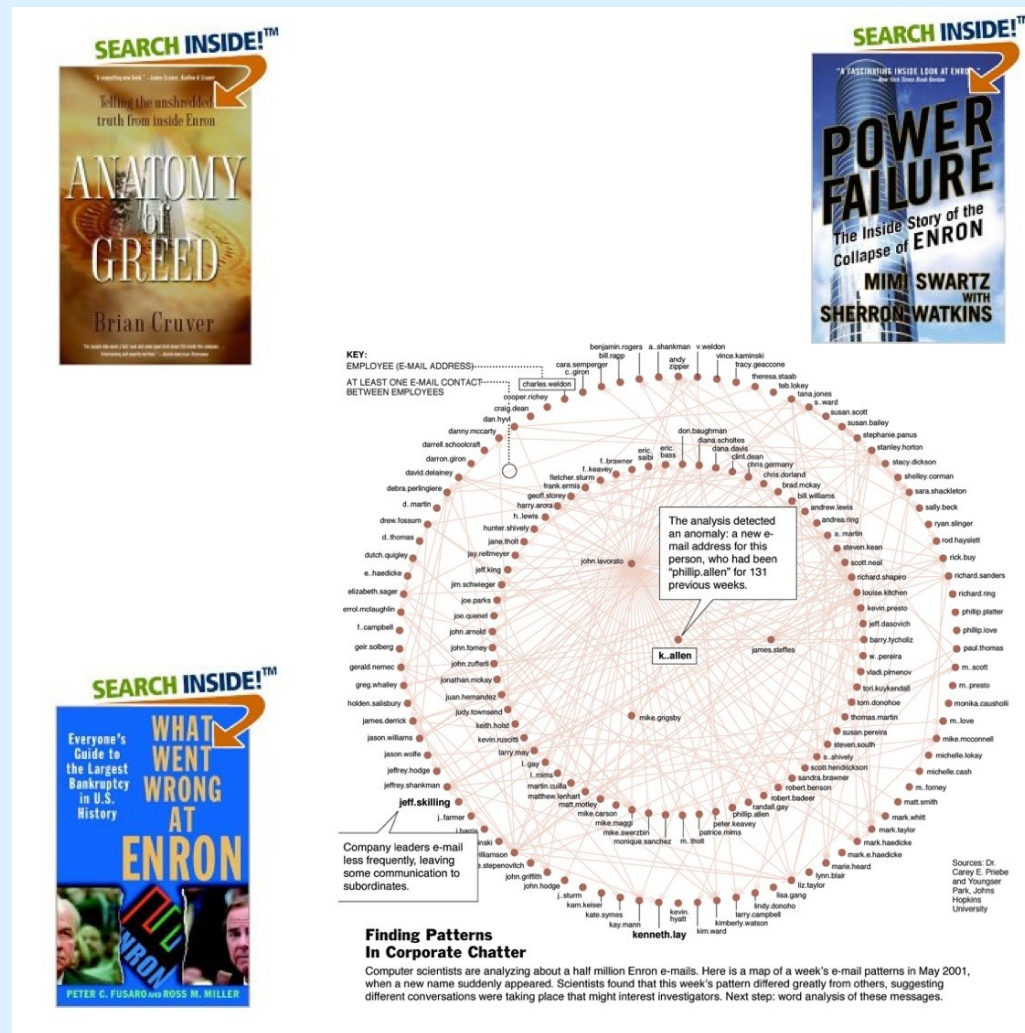
Supervised learning

Example 2: Find EVIL-DOERS



Enron Figure: The New York Times – Carey Priebe and Youngser Park, Johns Hopkins University

Example 2: Find EVIL-DOERS



Enron Figure: The New York Times – Carey Priebe and Youngser Park, Johns Hopkins University

Unsupervised learning

What is (Text) Data?

Object x_i has q measurements:

$$x_i = (x_{i1}, x_{i2}, \dots, x_{iq})^T \in \mathbb{R}^q$$

- * All n objects in the dataset can be expressed as an $n \times q$ data matrix
- * known as vector-space model

Examples:

- 1.) Text Mining: n documents, q weights or scores for particular words or phrases
- 2.) Image Analysis: n images, q pixel color or intensity values
- 3.) Computer Network Traffic: n application or protocol flows, q network traffic counts or scores
- 4.) DNA Expression Microarrays: n genes (nucleotide sequences), q cell samples

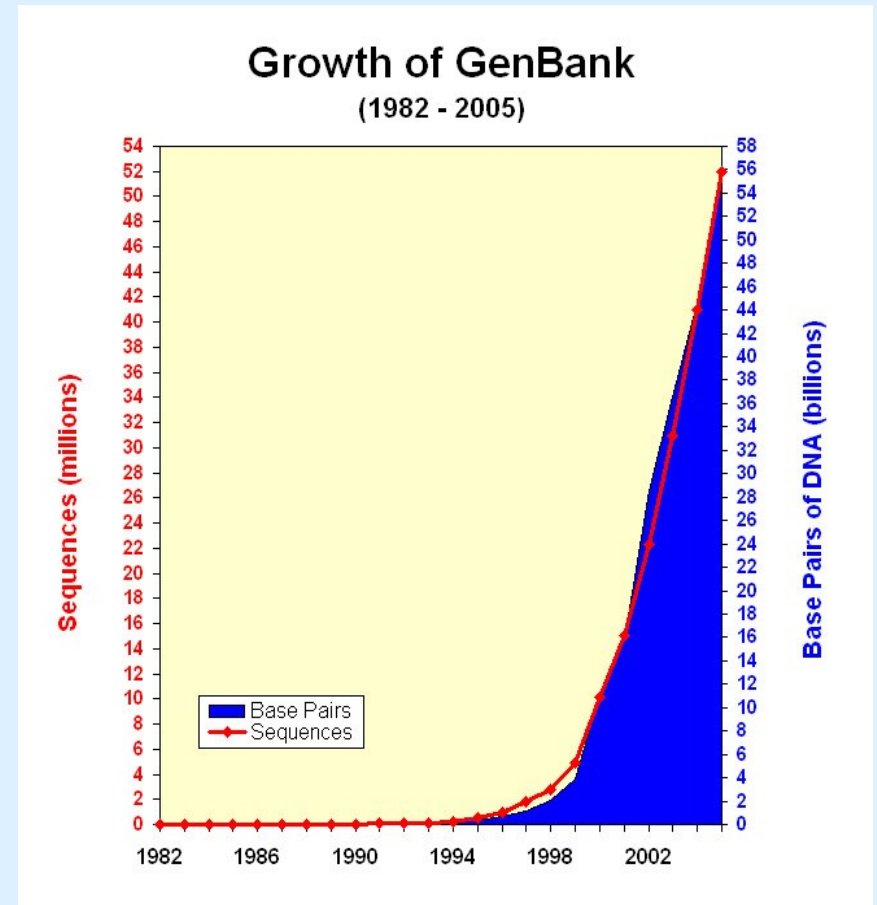
Growth of (Complex) Data

Amount of Data is Increasing

- * Figure: annotated collection of all available DNA sequences.
- * PUB MED: abstracts for 12 million research papers on life sciences topics. 40,000 new biomedical abstracts are added every month.
- * Similar growth of data in other fields: computer network traffic, text documents, images.

Complexity of Data is Increasing

- * n observations, q features/measurements/dimensions
- * Data that we need to analyze is high-dimensional: q large
- * In complex data, often $q \gg n$



Real-world Text Mining Issues

- * large databases
- * distance
- * high dimensionality
- * overfitting
- * missing/noisy data
- * global versus local structure
- * user involvement

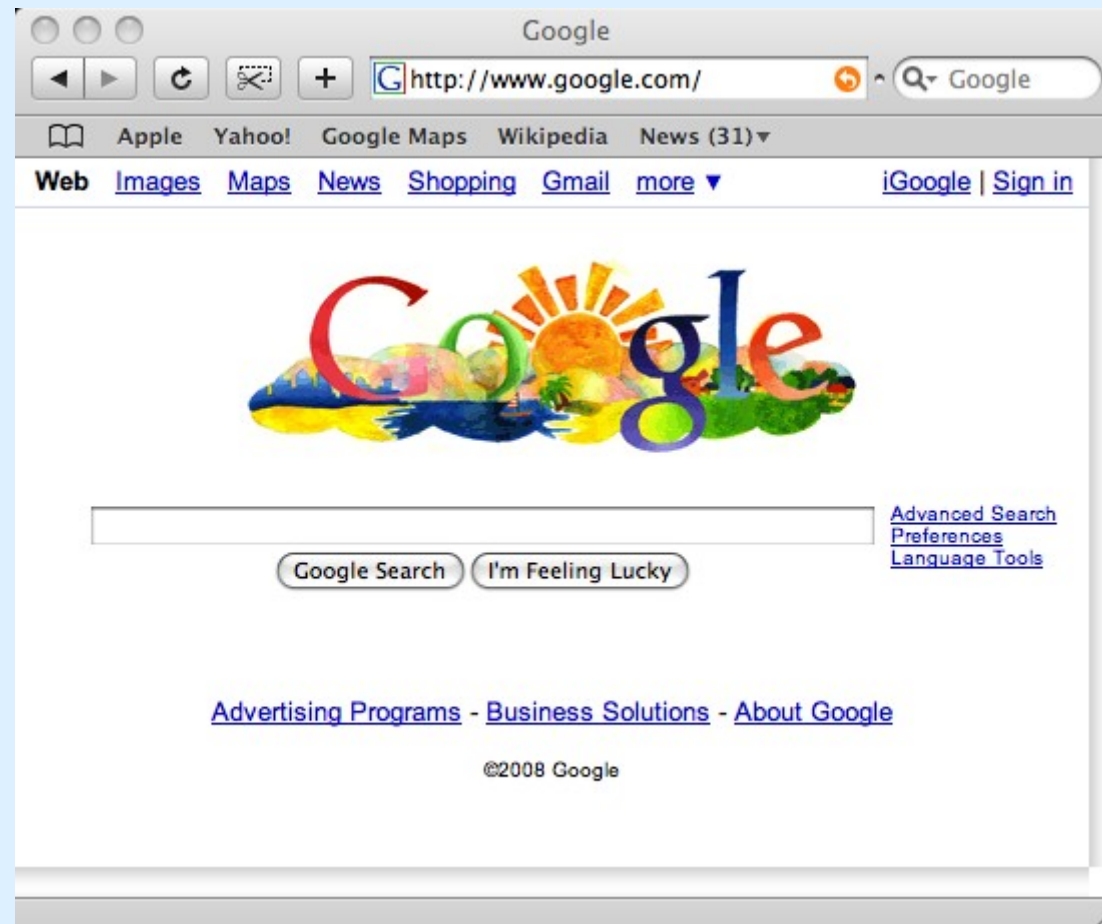
Real-world Text Mining Issues

- * large databases
- * distance
- * high dimensionality
- * overfitting
- * missing/noisy data
- * global versus local structure
- * user involvement ← ***Neglected!***

Non-linear synergy between quality of classifier and quality of user interaction

Non-linear synergy between quality of classifier and quality of user interaction

E.g.:



Interactive Text Mining with **Iterative Denoising**

The Iterative Denoising Methodology

In a nutshell, Iterative Denoising is a classification framework characterized by :

Processing a set of high-dimensional data; performing a local structure-preserving projection into a low-dimensional space; providing a visualization and interaction interface; partitioning and iteratively denoising.

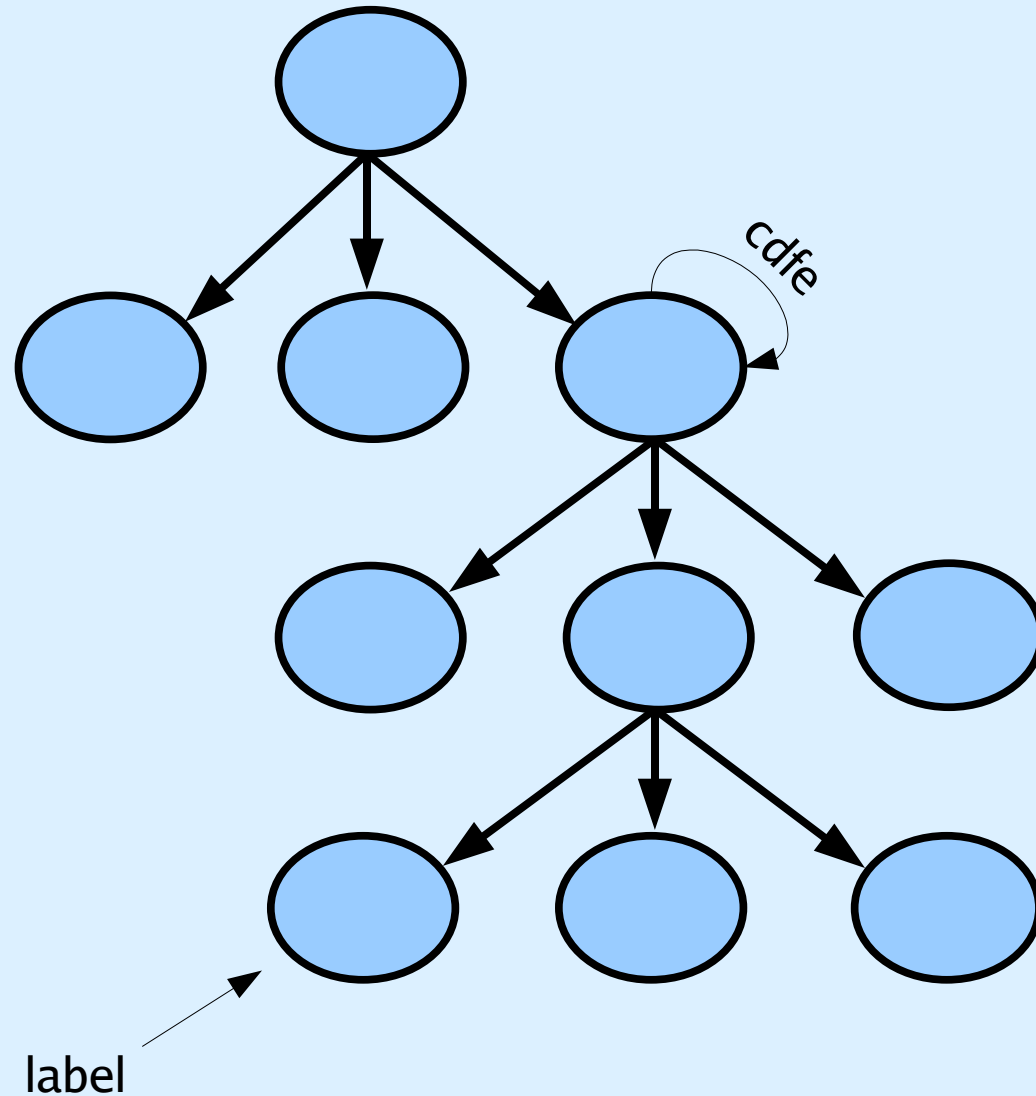
Motivated by:

Priebe, Marchette, Healy, 2004, “Integrated Sensing and Processing Decision Trees”, IEEE PAMI.

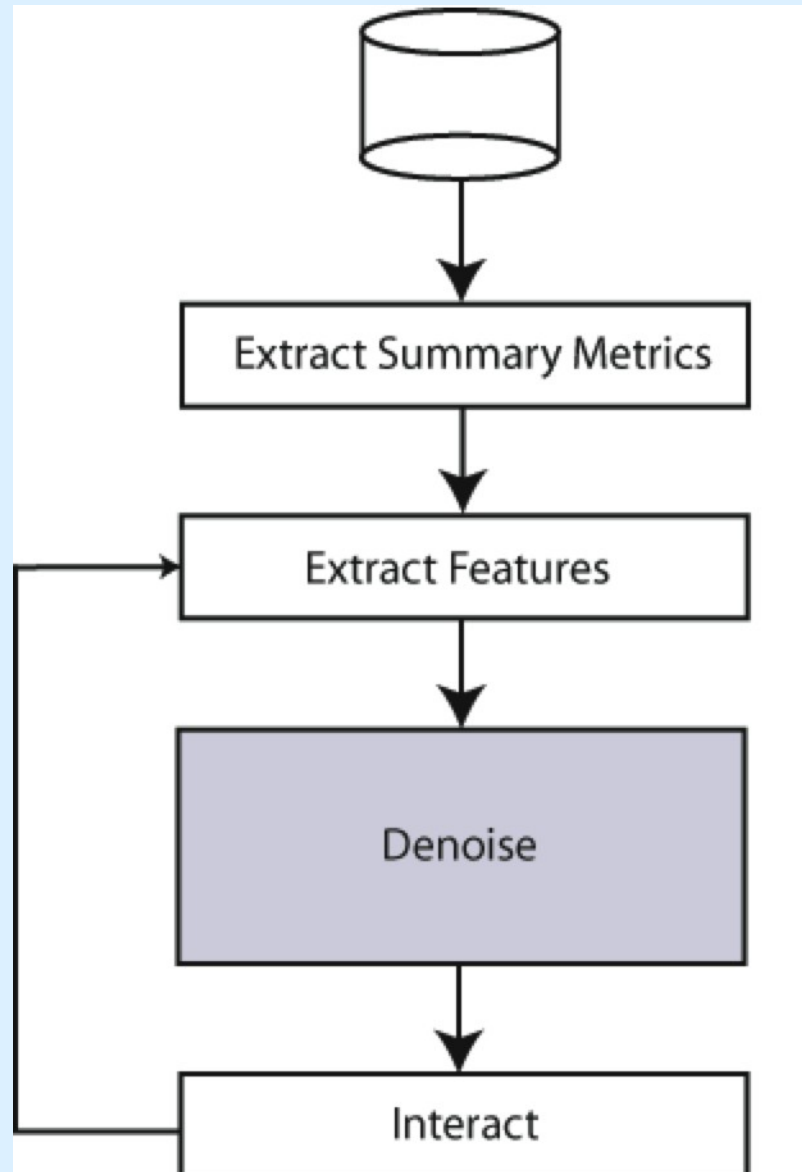
Priebe, et al., 2004, “Iterative Denoising for Cross-Corpus Discovery”, COMPSTAT.

The Concept: An Iterative Denoising Tree

Denoise



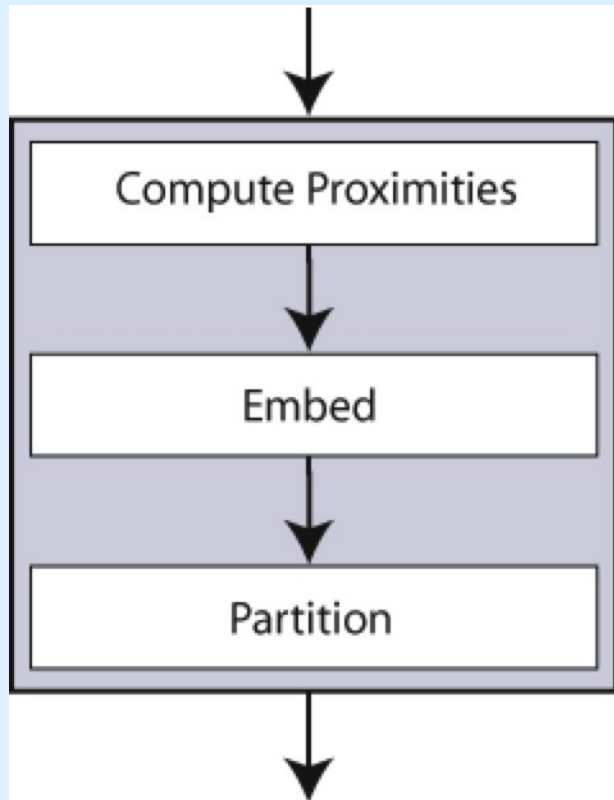
Iterative Denoising Framework



$$\Delta = \{\Delta_1, \dots, \Delta_n\} = \text{essentials}(\mathcal{C})$$

$$X_\Delta = \text{cdf}(\Delta)$$

Denoising Detail



A proximity metric

$$r_{ij} = \frac{\langle y_i, y_j \rangle}{\|y_i\| \|y_j\|}$$

A low dim space

$$F = \left[\frac{v_1}{\sqrt{\lambda_1}} \mid \cdots \mid \frac{v_d}{\sqrt{\lambda_d}} \right]$$

Clusters

$$W(C_1, \dots, C_k) = \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \bar{x}_i\|^2$$

Nonlinear version uses Laplacian Eigenmaps

Laplacian Eigenmaps

Nonlinear dimensionality reduction technique that distorts geometry in such a way that enhances some types of clustering

$L = D - A$ is large, sparse

L symmetric, positive semi-definite

$$0 < \lambda_1 \leq \lambda_2 \leq \dots \lambda_d$$

Corresponding d eigenvectors --> Fiedler Space

Eigenvectors corresponding to two smallest non-zero eigenvalues --> visualization

Stopping Criteria

Variety of methods can be used for stopping the tree build:

- * max tree height
- * min observations/node
- * (supervised): min node purity
- * (supervised): entropy or divergence

Here, we used min observations/node = $10d/k$

Interactive Text Mining with Iterative Denoising

Types of User Interactions

1. View Interaction

2. User-Guided Classification

3. Experience Imposition

Types of User Interactions

1. View Interaction

the user changes the given representations without affecting the object relationships

- * pan
- * zoom
- * object selection
- * link to data-space representation
- * switch between embedded views and tree views

Types of User Interactions

2. User-Guided Classification

the user affects the Iterative Denoising tree growth path

- * a form of selective discovery
- * shortens the data-processing to data-visualization interval
- * allows the user to change classification algorithm parameters at each tree node
- * facilitates data exploration and intuitive user hypothesis testing
- * useful step in online (streaming) analysis of massive data collections

Types of User Interactions

3. Experience Imposition

the user changes object relationships through qualitative inputs

- * what if these two objects were closer?
- * what if this object was more important?
- * what if this object was removed?
- * what if I now know an object's class label?

Interactive Text Mining with Iterative Denoising

A Text Document Example: a Science News Corpus

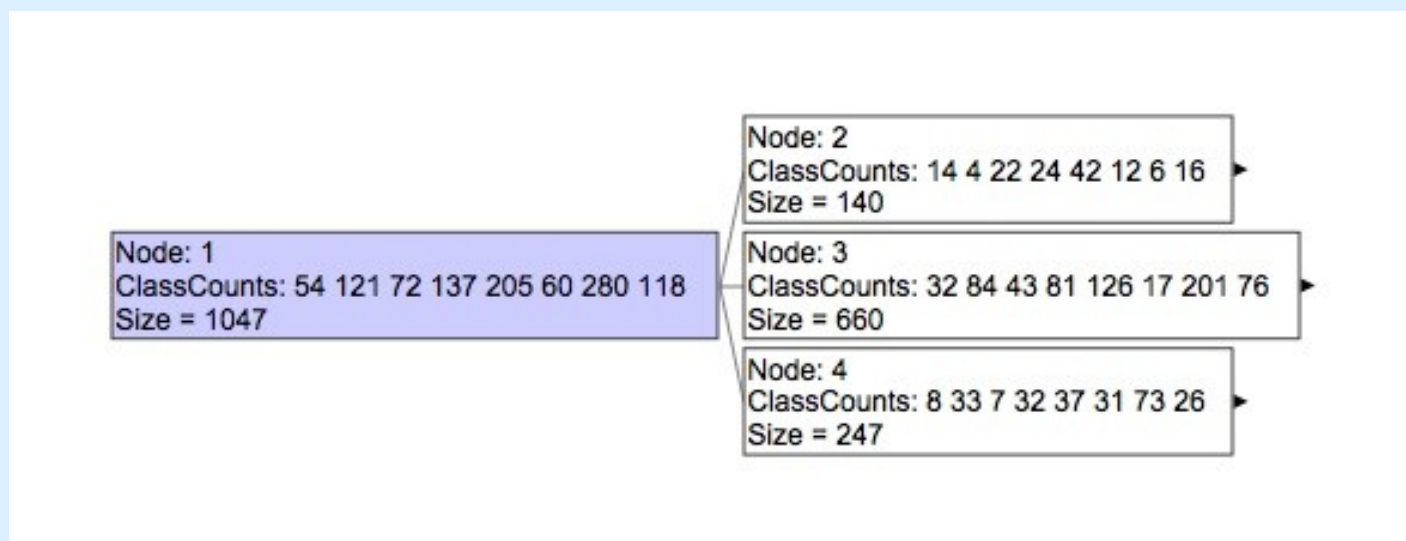


Class	Number of Documents
Anthropology	54
Astronomy	121
Behavioral Sciences	72
Earth Sciences	137
Life Sciences	205
Math & CS	60
Medicine	280
Physics	118

Table 1: Science News corpus.

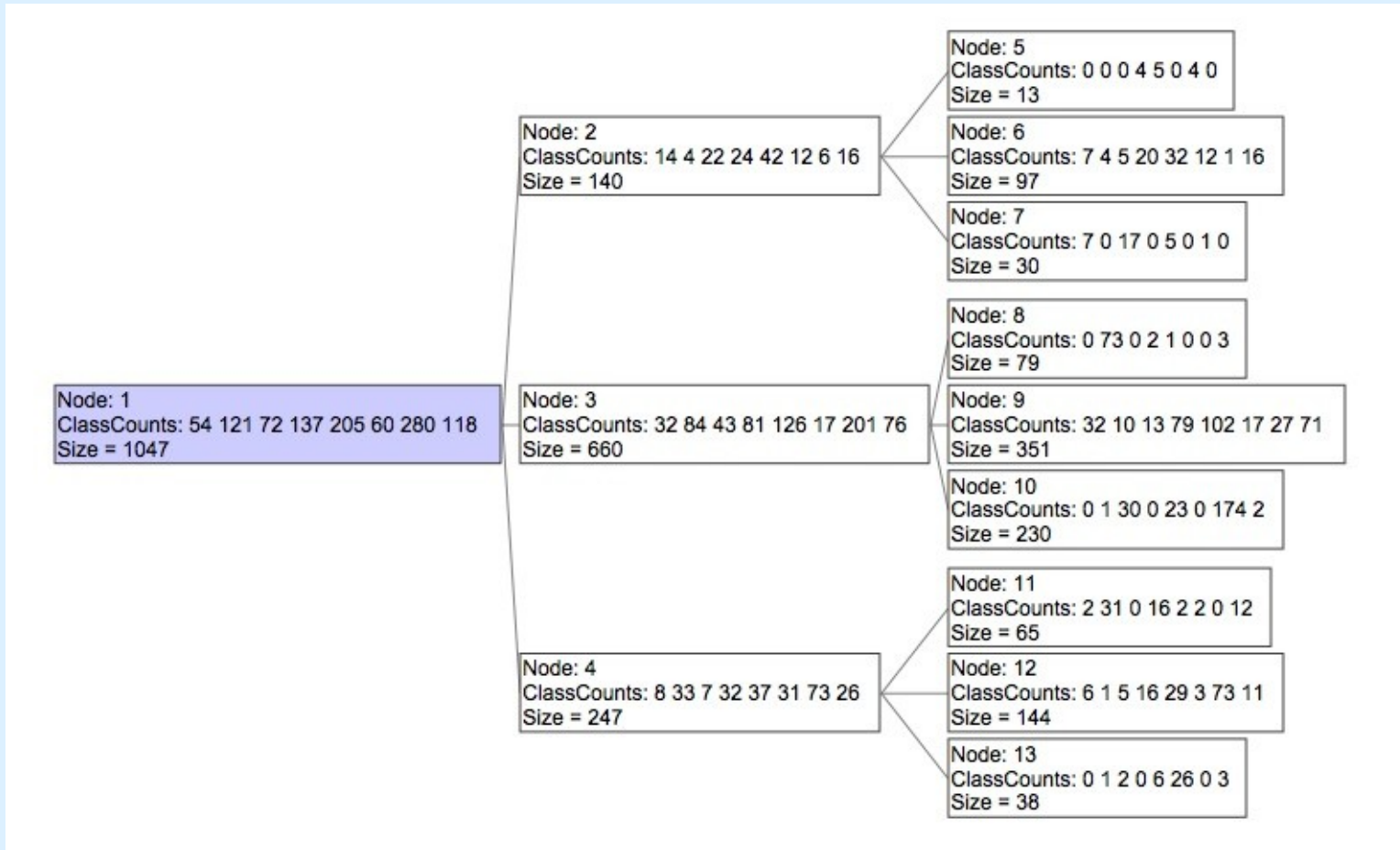
$n = 1047$ documents
 $q = 32130$ words (ngrams)

Science News corpus: a clustering hierarchy



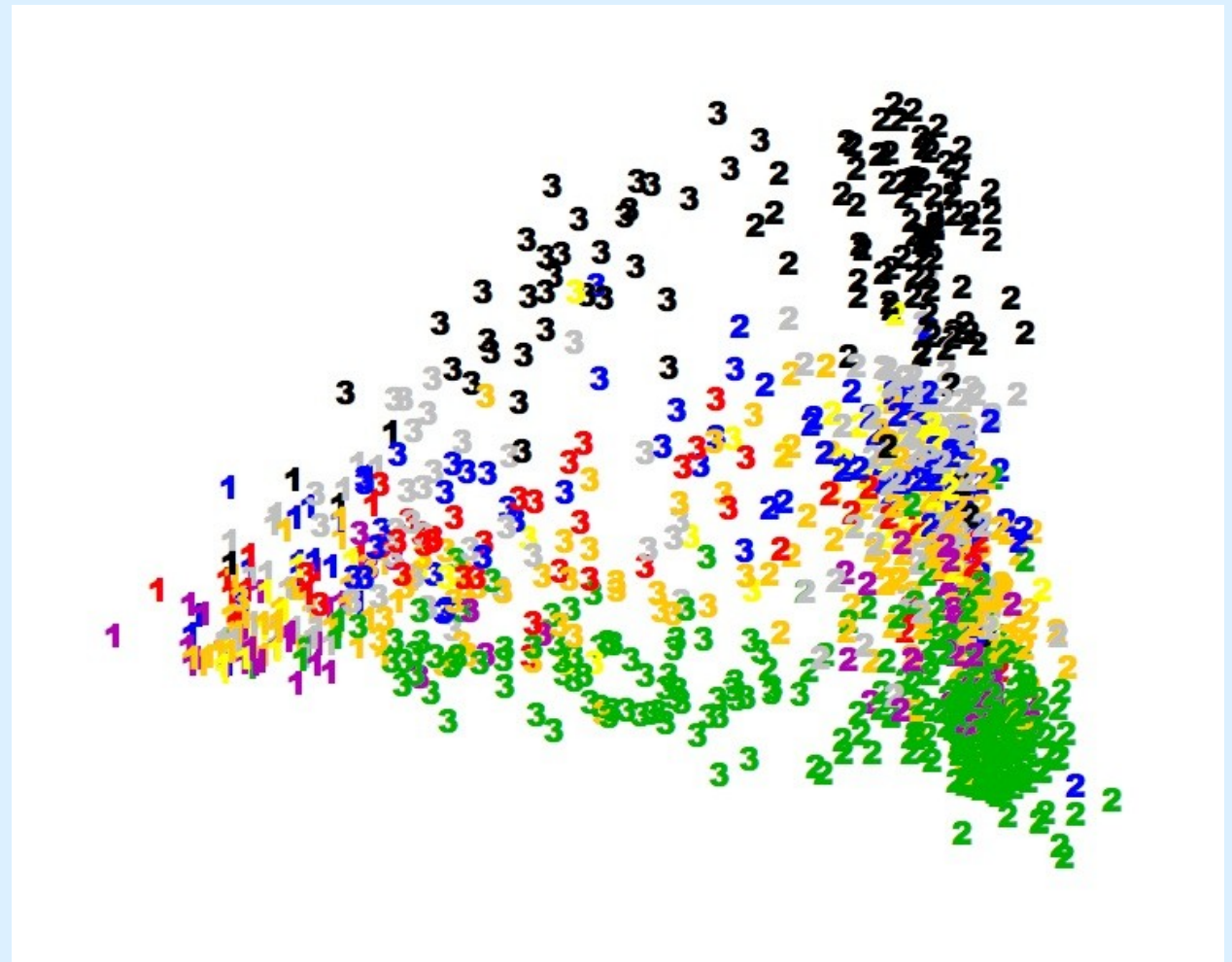
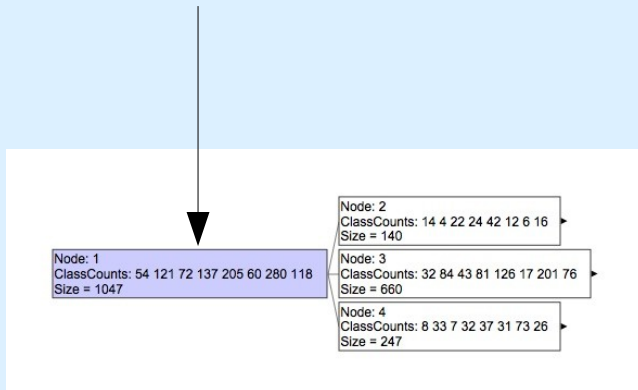
1. Anthropology
2. Astronomy
3. Behavioral Sciences
4. Earth Sciences
5. Life Sciences
6. Math & CS
7. Medicine
8. Physics

Science News corpus: a clustering hierarchy



1. Anthropology
2. Astronomy
3. Behavioral Sciences
4. Earth Sciences
5. Life Sciences
6. Math & CS
7. Medicine
8. Physics

Science News corpus: Fiedler Space projection



Anthropology: yellow
Astronomy: black
Behavioral Sciences: magenta
Earth Sciences: lightGray
Life Sciences: orange
Math & CS: red
Medicine: green
Physics: blue

20-Newsgroups

* Collection of newsgroup documents partitioned into 20 different newsgroups.
--> <http://people.csail.mit.edu/jrennie/20Newsgroups/>

* As a simple test, we chose three roughly disparate groups, and two groups on the same topic but with differing viewpoints.

- 1.) rec.autos
- 2.) talk.politics.misc
- 3.) comp.graphics

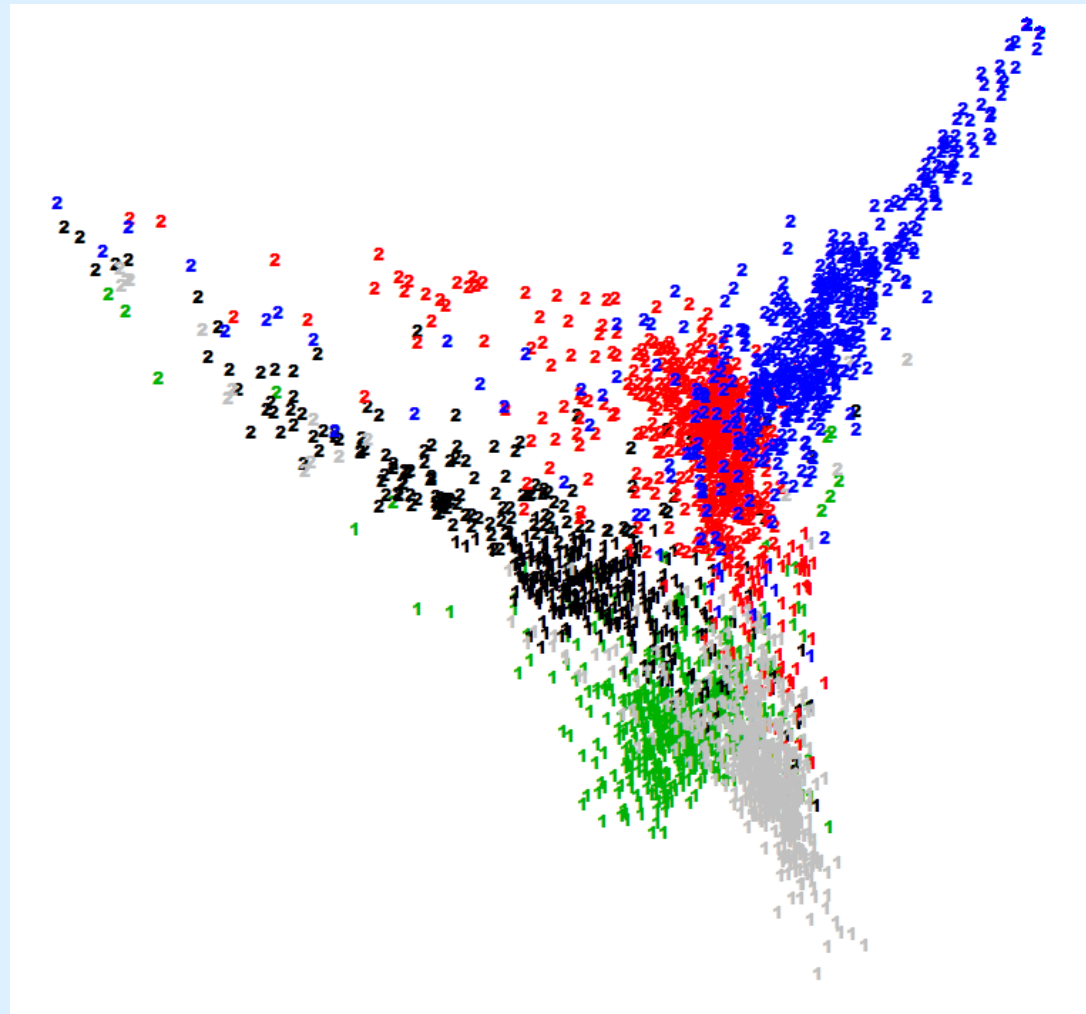
- 4.) soc.religion.christian
- 5.) alt.athiesm

n = 2722 documents, with 25887 unique monograms

* NB: we did not do typical text pre-processing tasks such as stop-word removal, removal of infrequent words, etc.

20-Newsgroups: Iterative Denoising

green	rec.autos
black	talk.politics.misc
red	soc.religion.christian
gray	comp.graphics
blue	alt.athiesm



Interactive Text Mining with Iterative Denoising

Demo

<http://www.people.vcu.edu/~kegiles/>

Kendall Giles

Assistant Professor

[Department of Statistical Sciences and Operations
Research](#)

[College of Humanities and Sciences
Virginia Commonwealth University](#)

Associate Staff Scientist

[Human Language Technology Center for Excellence
\(HLTCOE\)
Johns Hopkins University](#)

Contact Information

Email: first two letters of first name + last name +
@vcu.edu
Office: Oliver 2061
Mail: PO Box 843083
1001 West Main Street
Richmond, VA 23284

