Basics of Knowledge Discovery Engines

Kendall Giles

Department of Computer Science
Johns Hopkins University
and
Department of Statistical Sciences and Operations Research
Virginia Commonwealth University

UCLA IPAM Tutorial, September 11, 2007

Outline

Outline

What the Customer Wants

A mysterious figure...

who wants data turned into knowledge...

What the Customer Wants

A mysterious figure...

who wants data turned into knowledge...

What the Customer Wants

A mysterious figure...

who wants data turned into knowledge...

What the Users Need

Explorers of data...

who need help with the deluge...

What the Users Need

Explorers of data...

who need help with the deluge...

What the Users Need

Explorers of data...

who need help with the deluge...

Outline

How to Solve Problems

The Feynman Problem-Solving Algorithm:

1. write down the problem;
2. think very hard;
3. write down the answer.

attributed to Murray Gell-Mann

How to Solve Problems

The Feynman Problem-Solving Algorithm:

1. write down the problem;
2. think very hard;
3. write down the answer.

attributed to Murray Gell-Mann

How to Solve Problems

The Feynman Problem-Solving Algorithm:

1. write down the problem;
2. think very hard;
3. write down the answer.

attributed to Murray Gell-Mann

How to Solve Problems

The Feynman Problem-Solving Algorithm:

1. write down the problem;
2. think very hard;
3. write down the answer.

attributed to Murray Gell-Mann

How to Discover Knowledge

The Knowledge Discovery Algorithm:

1. collect some data;
2. look at it;
3. collect the knowledge.

How to Discover Knowledge

The Knowledge Discovery Algorithm:

1. collect some data;
2. look at it;
3. collect the knowledge.

How to Discover Knowledge

The Knowledge Discovery Algorithm:

1. collect some data;
2. look at it;
3. collect the knowledge.

How to Discover Knowledge

The Knowledge Discovery Algorithm:

1. collect some data;
2. look at it;
3. collect the knowledge.

Outline

What is Knowledge Discovery? 1

Goes by various names:

- data mining
- knowledge discovery
- machine learning
- pattern recognition
- statistical learning

What is Knowledge Discovery? 1

Goes by various names:

- data mining
- knowledge discovery
- machine learning
- pattern recognition
- statistical learning

What is Knowledge Discovery? 2

Is interdisciplinary:

- computational statistics
- machine learning
- visualization
- high-performance computing
- data storage systems
- algorithms
- operations research
- bioinformatics
- information retrieval

What is Knowledge Discovery? 2

Is interdisciplinary:

- computational statistics
- machine learning
- visualization
- high-performance computing
- data storage systems
- algorithms
- operations research
- bioinformatics
- information retrieval

What is Knowledge Discovery? 3

Can be defined as:

the process of extracting previously unknown and potentially useful patterns inherent in data

What is a pattern?

an expression of some subset of the data or a model of the subset, or a high-level description of some subset of the data

What is Knowledge Discovery? 3

Can be defined as:

the process of extracting previously unknown and potentially useful patterns inherent in data

What is a pattern?

an expression of some subset of the data or a model of the subset, or a high-level description of some subset of the data

What is Knowledge Discovery? 3

Can be defined as:

> *the process of extracting previously unknown and potentially useful patterns inherent in data*

What is a pattern?

> *an expression of some subset of the data or a model of the subset, or a high-level description of some subset of the data*

What is Knowledge Discovery? 3

Can be defined as:

> *the process of extracting previously unknown and potentially useful patterns inherent in data*

What is a pattern?

> *an expression of some subset of the data or a model of the subset, or a high-level description of some subset of the data*

What is Knowledge Discovery? 4

Note:

clever algorithms are good, but clever algorithms that can be implemented for a user are even better

classic tradoff between: speed, accuracy, cost

What is Knowledge Discovery? 4

Note:

clever algorithms are good, but clever algorithms that can be implemented for a user are even better

classic tradoff between: speed, accuracy, cost

What is Knowledge Discovery? 4

Note:

clever algorithms are good, but clever algorithms that can be implemented for a user are even better

classic tradoff between: speed, accuracy, cost

What is Data?

Object x_i has q measurements:

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

- 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

What is Data?

Object x_i has q measurements:

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

- 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

Many (Statistical) Approaches

class-conditional densities:

- known
  - Bayes Decision Theory
- unknown
  - supervised
    - parametric
    - nonparametric
  - unsupervised
    - parametric
    - nonparametric

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 foroperative 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 effortsin 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 countriesstart to buy gold for their reserves.But what is important is that a start has beenmade 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 2007Lisbon, 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 Fern�ndez Ard�vol, Universitat Oberta de Catalunya (UOC), Barcelona, Spain

* Conference Background and GoalsThe 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 contributes 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 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 foroperative 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 effortsin 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 countriesstart to buy gold for their reserves.But what is important is that a start has beenmade 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 2007Lisbon, 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)

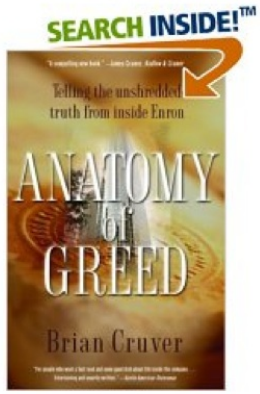
* Keynote Speaker (confirmed)Professor Mireia Fern�ndez Ard�vol, Universitat Oberta de Catalunya (UOC), Barcelona, Spain

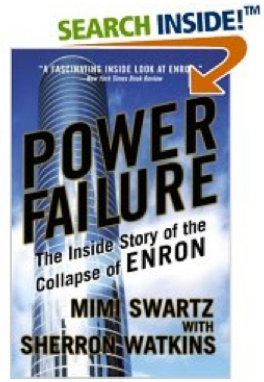
* Conference Background and GoalsThe IADIS e-Society 2007 conference aims to address the main issues of concern within the Information Society. This conference covers both the technical as well as the non-technical aspects of the Information Society. Broad areas of interest are eGovernment / eGovernance,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

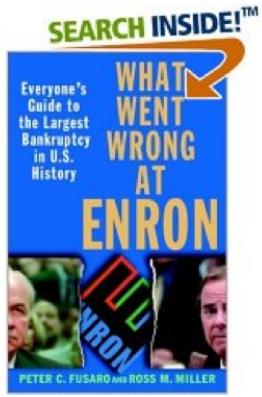
* Format of the Conference The conference will comprise of invited talks and oral presentations. The proceedings of the conference will be published in the form of a book and CD-ROM with ISBN, and will be available also in the IADIS Digital Library (accessible on-line). The best paper authors will be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

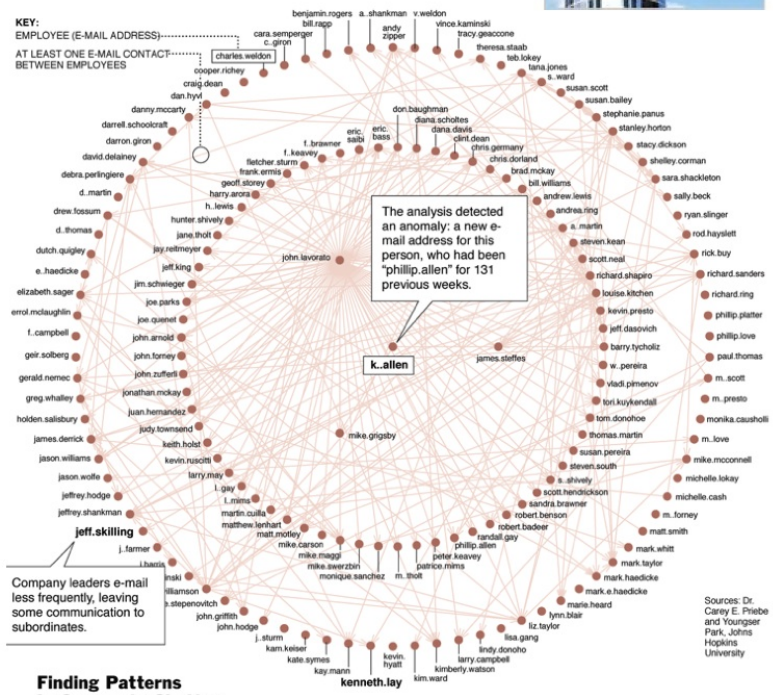
supervised learning

Example 2: Find EVIL-DOERS



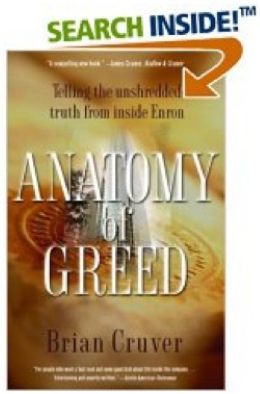


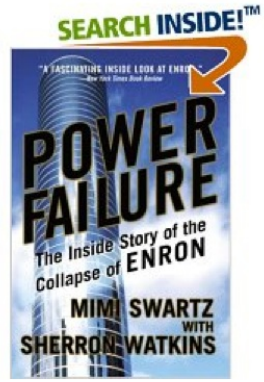


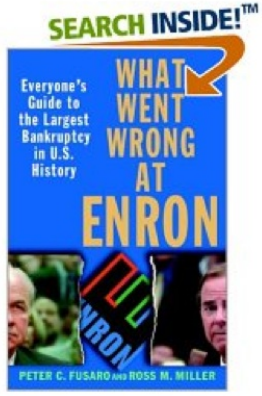


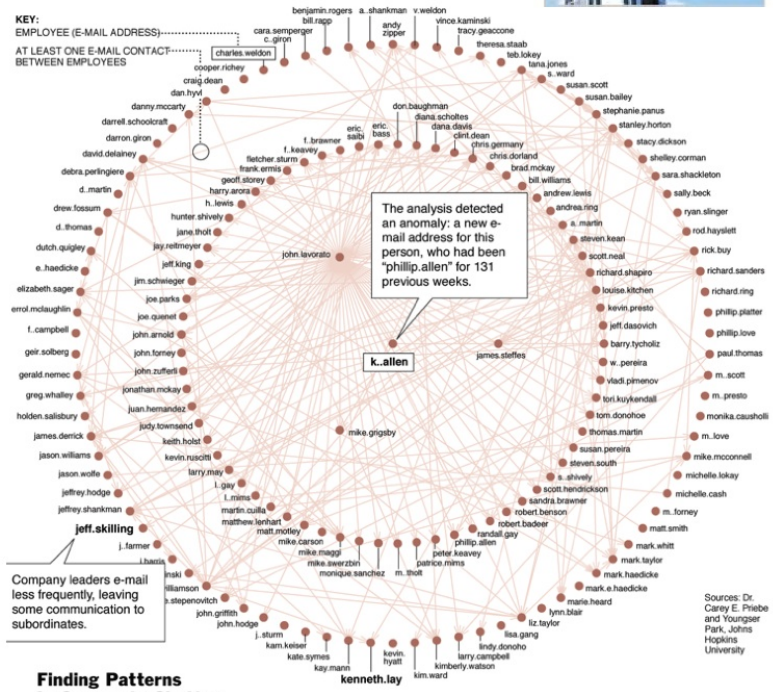
Enron Figure: The New York Times; Carey Priebe and Youngser Park, Johns Hopkins University

Example 2: Find EVIL-DOERS









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

unsupervised learning

Method Components

1. model representation
2. model evaluation
3. parameter/model search

Implementation Issues

- large databases
- distance
- high dimensionality
- overfitting
- missing/noisy data
- local patterns (as opposed to global patterns)
- user involvement in the search

Outline

Process

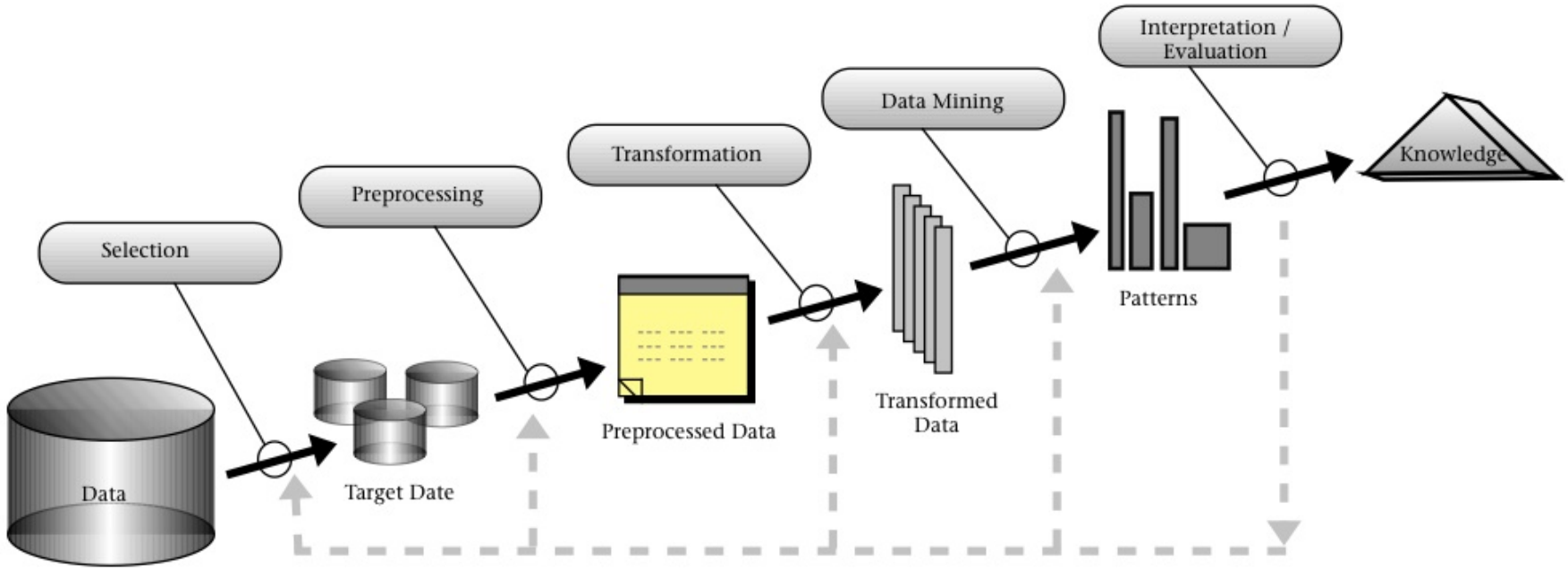


Figure: Fayyad, et al.

Outline

Iterative Denoising Methodology

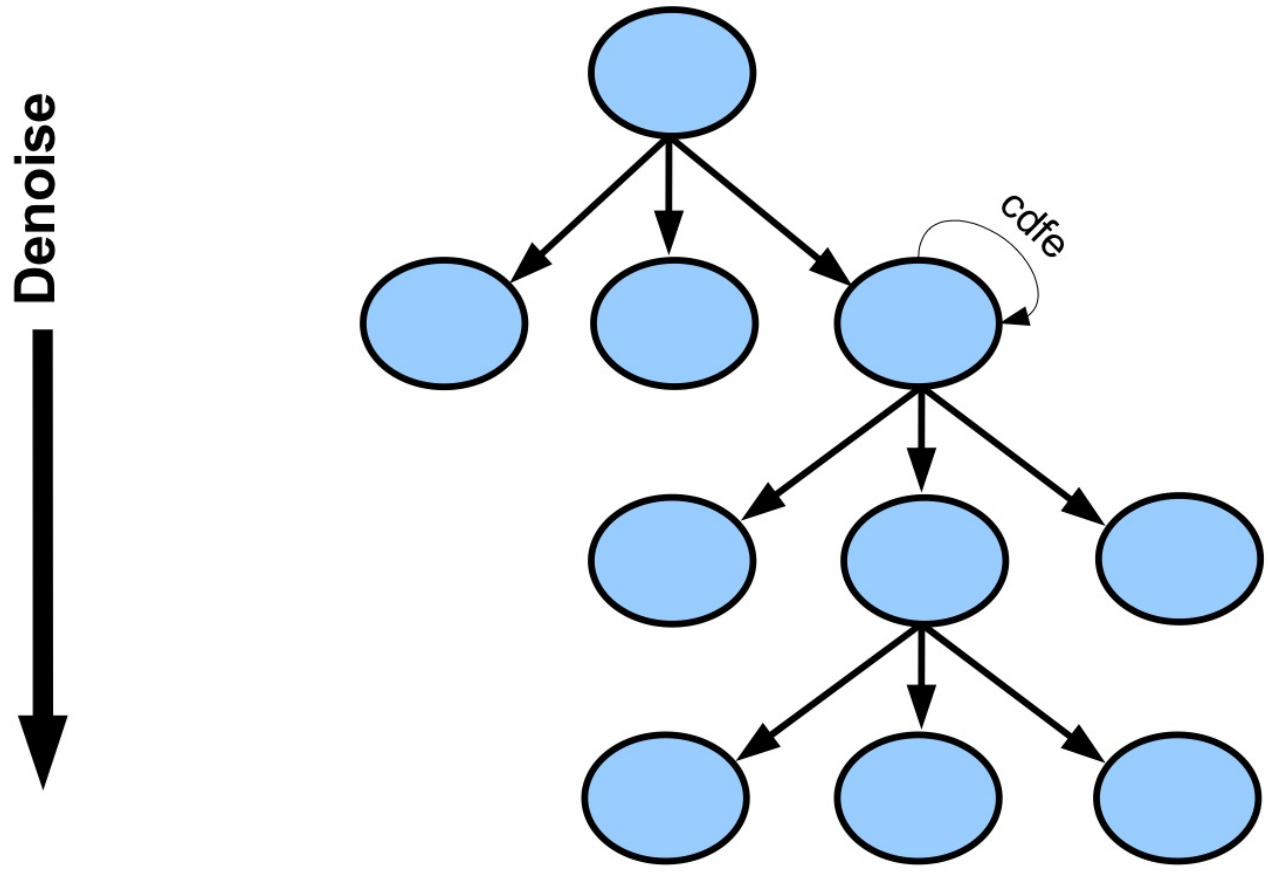
In a nutshell:

> *Process a set of high-dimensional data; perform a local structure-preserving projection into a low-dimensional space; provide a visualization and interaction interface; partition and iteratively denoise.*

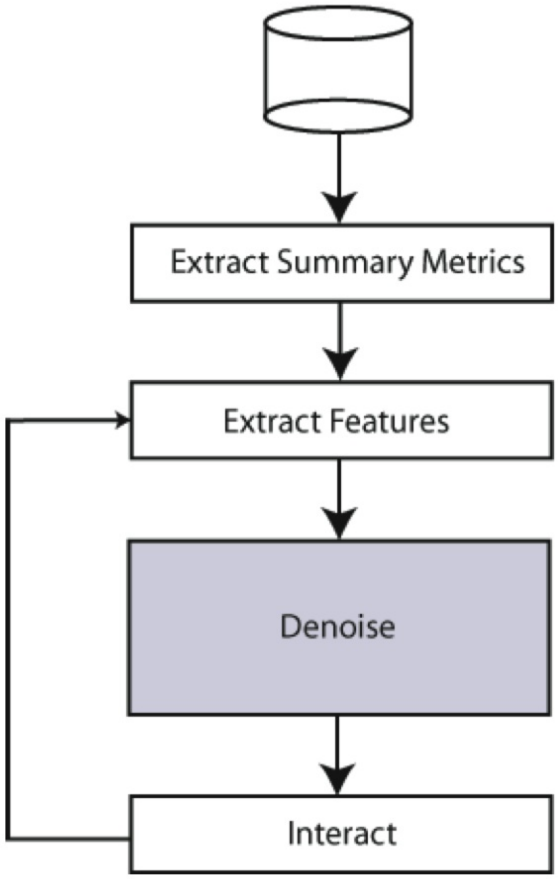
Motivated by:

- Priebe, Marchette, Healy, 2004, "Integrated Sensing and Processing Decision Trees," *IEEE PAMI*.
- Priebe, et al., 2004, "Iterative Denoising for Cross-Corpus Discovery," *COMPSTAT*.

An Iterative Denoising Tree



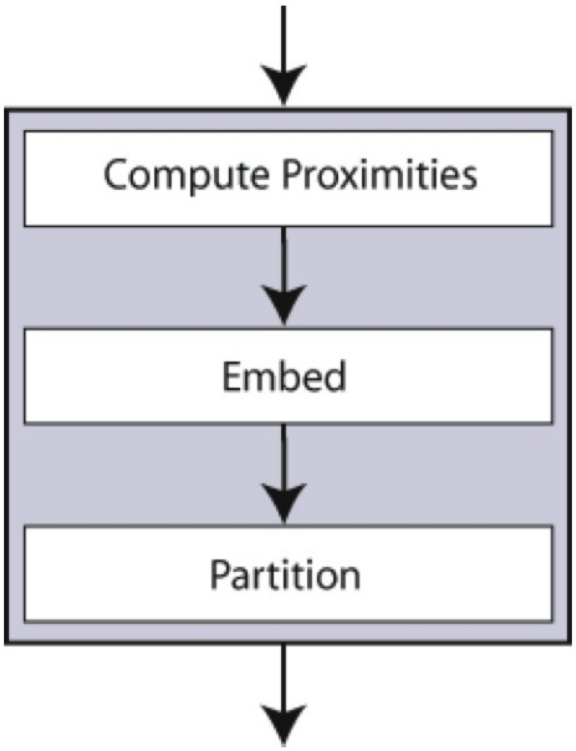
Iterative Denoising Framework



$$\Delta = \{\Delta_1, ..., \Delta_n\} = \texttt{essentials}(\mathcal{C})$$

$$X_\Delta = \texttt{cdfe}(\Delta)$$

Denoising Detail



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

A low dim space $$F = \left[\frac{v_1}{\sqrt{\lambda_1}} \middle| \cdots \middle| \frac{v_d}{\sqrt{\lambda_d}} \right]$$

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

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** is symmetric, positive semi-definite
- $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$
- corresponding d eigenvectors –> Fiedler Space
- eigenvectors corresponding to two smallest non-zero eigenvalues –> visualization

Outline

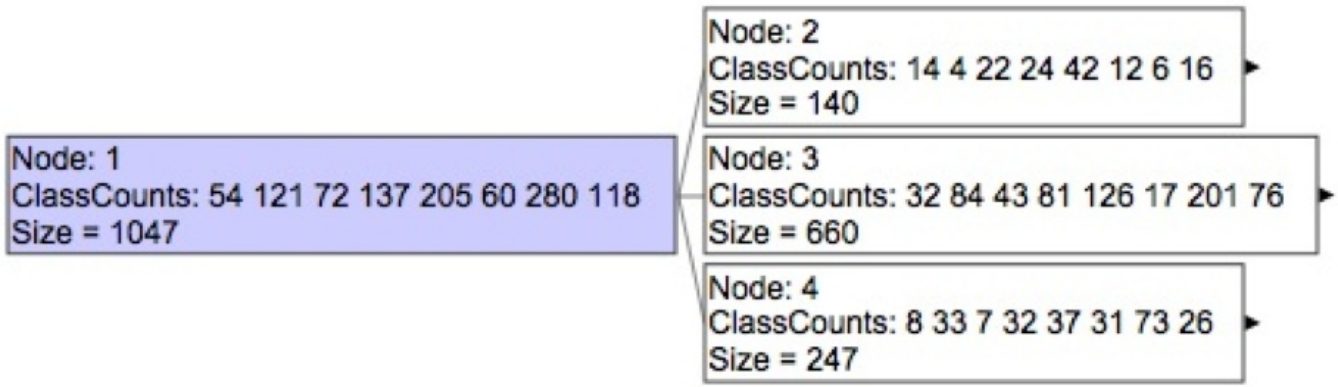
A Science News Corpus

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.

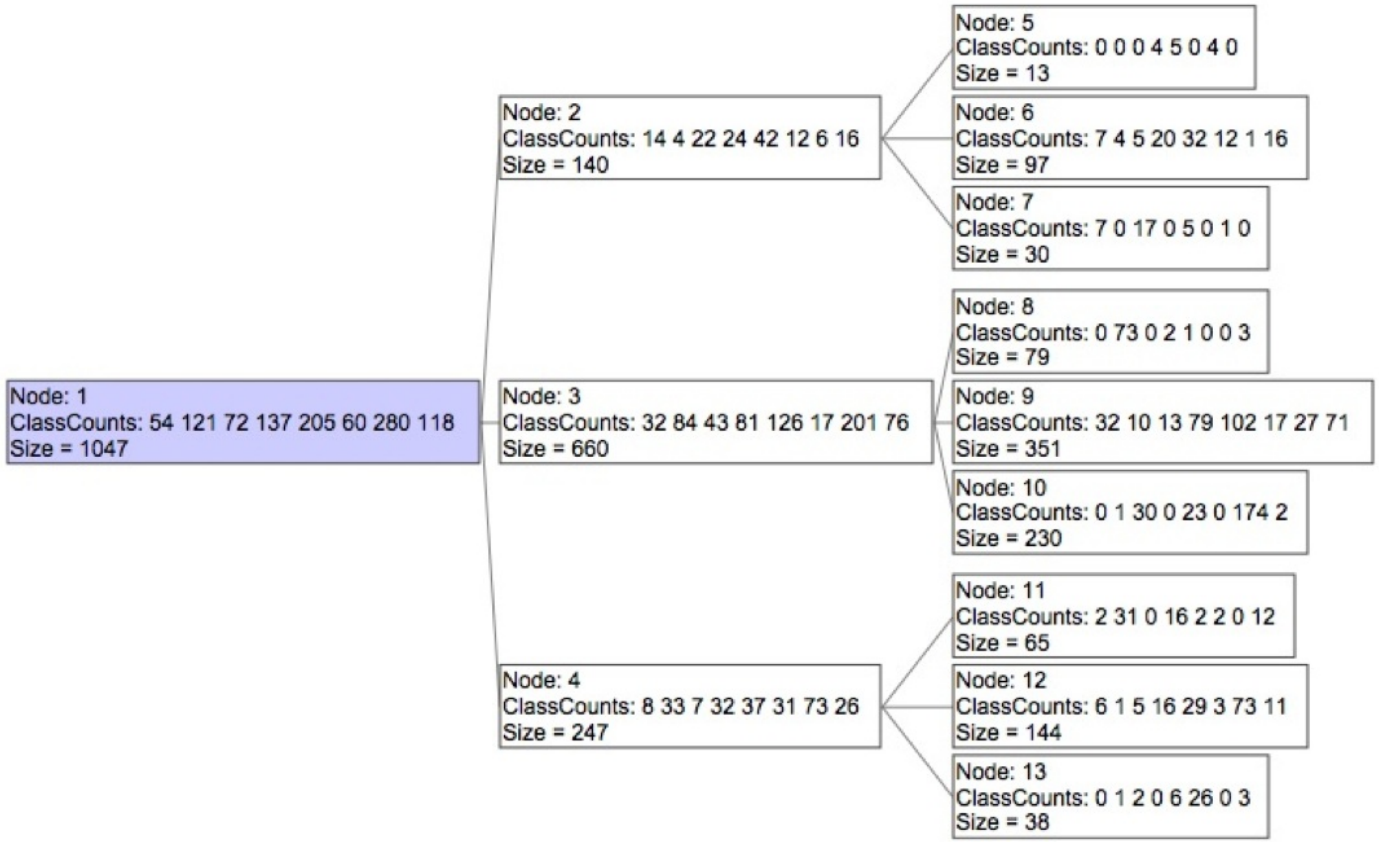
n = 1047 documents
q = 32130 words (ngrams)

A Clustering Hierarchy 1



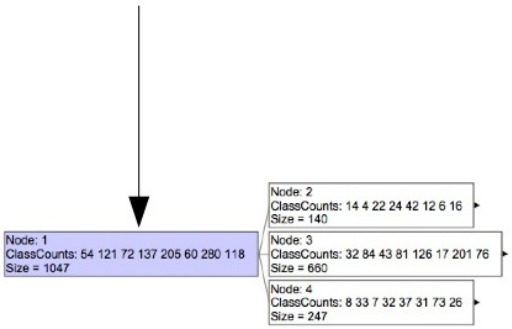
1. Anthropology
2. Astronomy
3. Behavioral Sciences
4. Earth Sciences
5. Life Sciences
6. Math & CS
7. Medicine
8. Physics

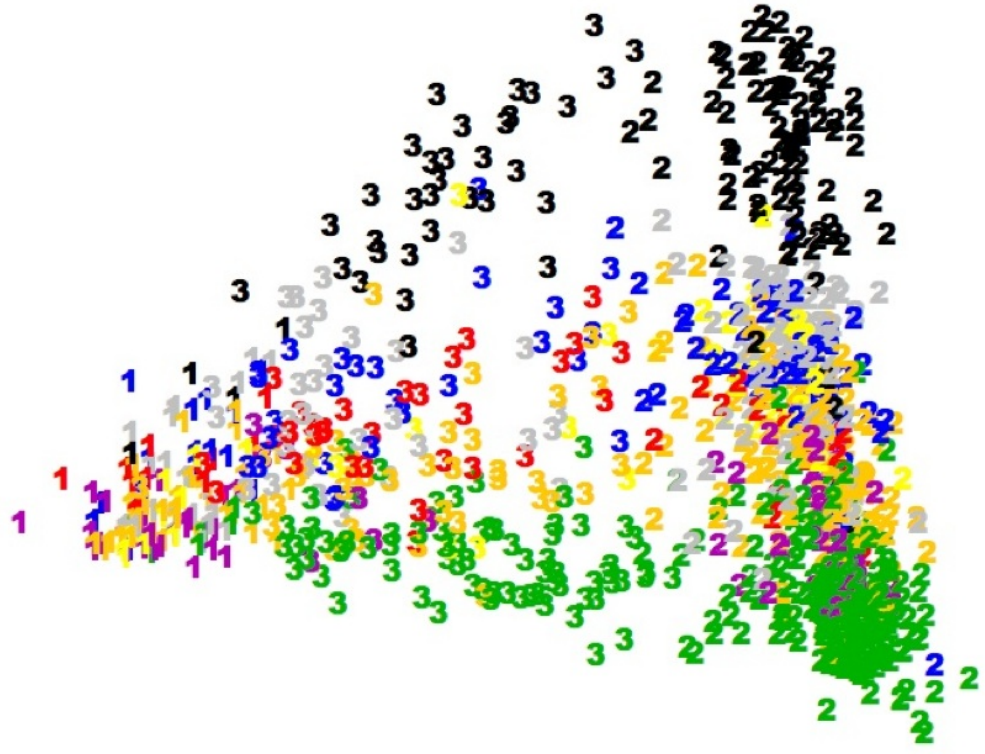
A Clustering Hierarchy 2



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

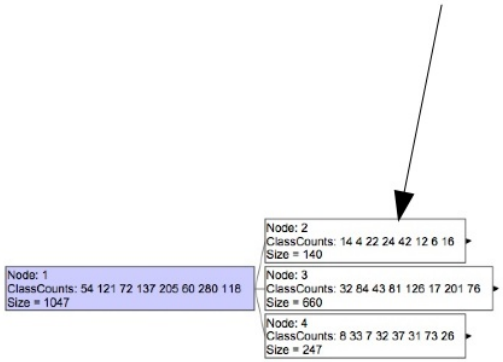
Fiedler Space Projection 1





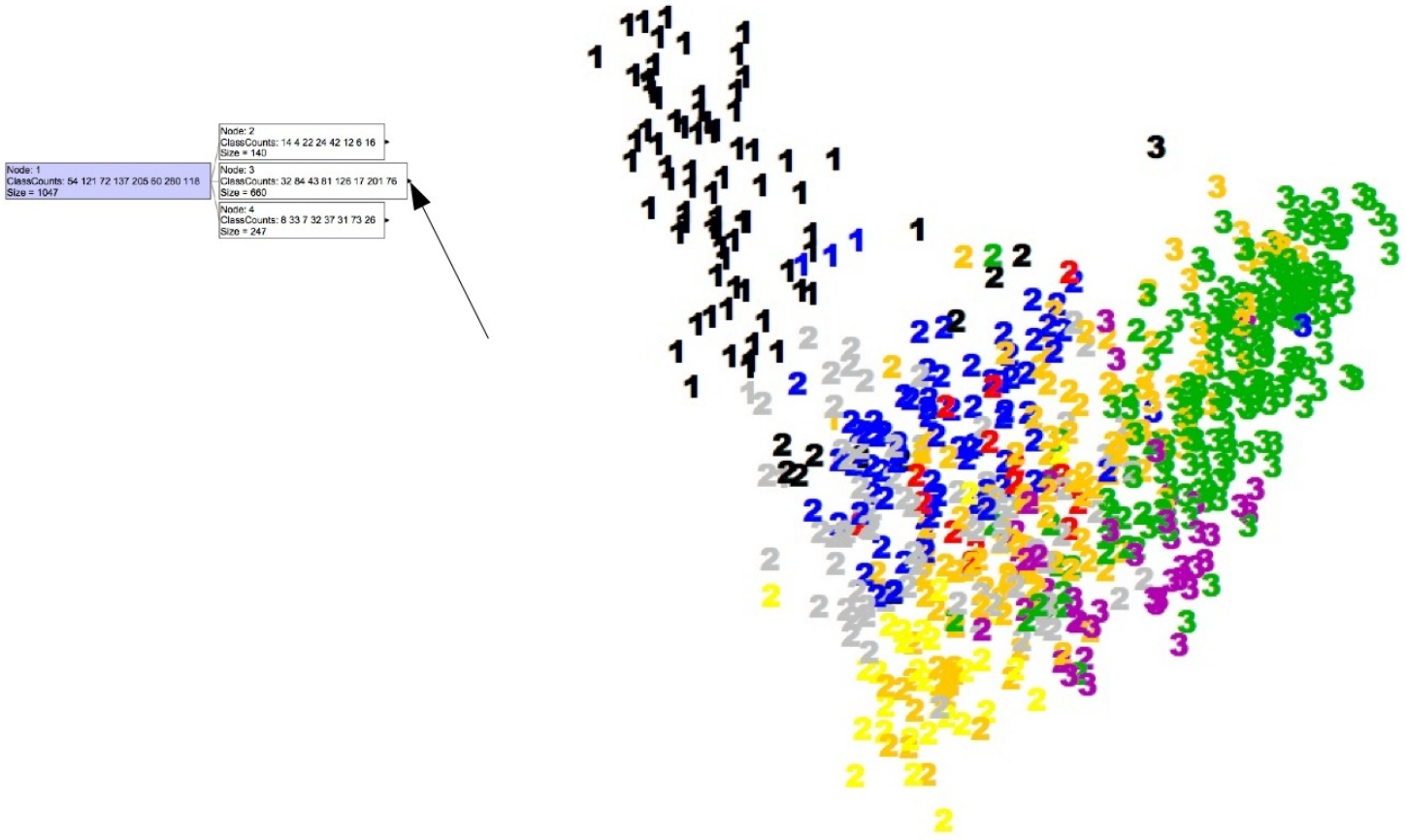
1. Anthropology: yellow
2. Astronomy: black
3. Behavioral Sciences: magenta
4. Earth Sciences: lightGray
5. Life Sciences: orange
6. Math & CS: red
7. Medicine: green
8. Physics: blue

Fiedler Space Projection 2



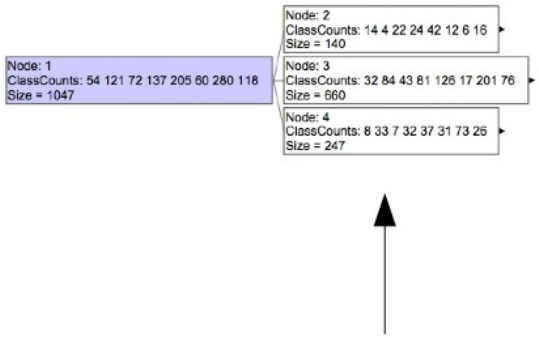
1. Anthropology: yellow
2. Astronomy: black
3. Behavioral Sciences: magenta
4. Earth Sciences: lightGray
5. Life Sciences: orange
6. Math & CS: red
7. Medicine: green
8. Physics: blue

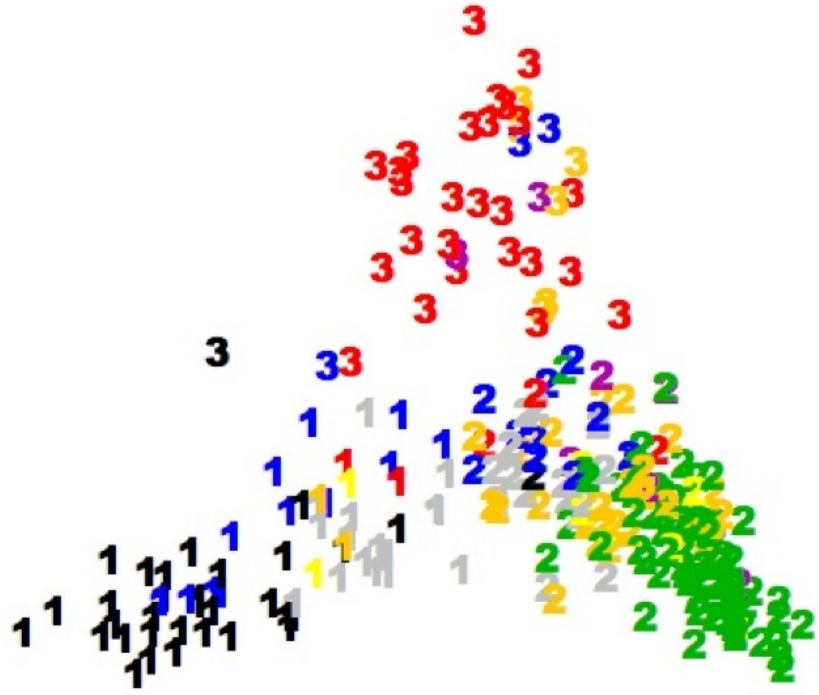
Fiedler Space Projection 3



1. Anthropology: yellow
2. Astronomy: black
3. Behavioral Sciences: magenta
4. Earth Sciences: lightGray
5. Life Sciences: orange
6. Math & CS: red
7. Medicine: green
8. Physics: blue

Fiedler Space Projection 4

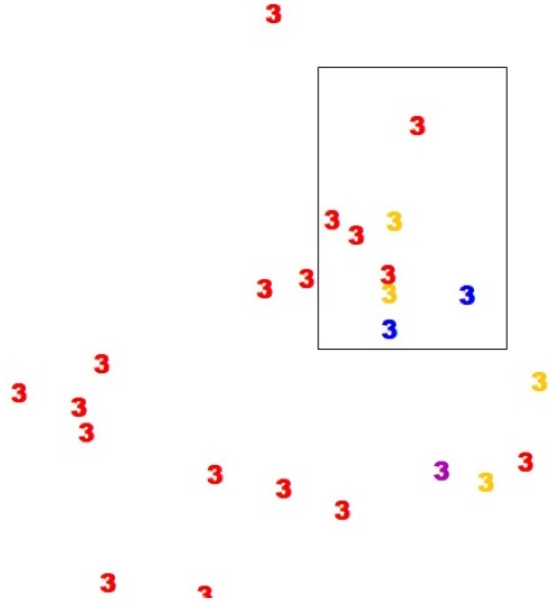


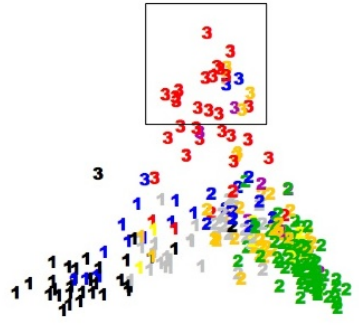


1. Anthropology: yellow
2. Astronomy: black
3. Behavioral Sciences: magenta
4. Earth Sciences: lightGray
5. Life Sciences: orange
6. Math & CS: red
7. Medicine: green
8. Physics: blue

Interesting Grouping

1. “Math enthusiast wins Science Talent Search”
2. “Message in DNA tops Science Talent Search”
3. “Chinks in Digital Armor: exploiting faults to break smart-card cryptosystems”
4. “Motor City hosts top science fair winners”
5. “Science Talent Search winners shine bright”
6. “Neutrinos to buckyballs: 10 talents tower”
7. “Logic in the Blocks: simple puzzles can give computers an unexpected workout”
8. “How to trick other people's computers into solving your math problems”





1. Anthropology: yellow
2. Astronomy: black
3. Behavioral Sciences: magenta
4. Earth Sciences: lightGray
5. Life Sciences: orange
6. Math & CS: red
7. Medicine: green
8. Physics: blue

Thank you.

Basics of Knowledge Discovery Engines

Kendall Giles

kgiles@cs.jhu.edu

www.kendallgiles.com

JOHNS HOPKINS UNIVERSITY

Institute for Pure & Applied Mathematics
IPAM
University of California, Los Angeles

VCU

For Further Reading:


K. Giles, M. Trosset, D. Marchette, C. Priebe.
Iterative Denoising.
Computational Statistics, Accepted for publication. 2007.


U. Fayyad, G. Piatetsky-Shapiro, P. Smyth.
From Data Mining to Knowledge Discovery in Databases.
AI Magazine, 17:37–54, 1996.


A. Jain, R. Duin, J. Mao.
Statistical Pattern Recognition: A Review.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1):4–37. 1998.