Data Mining
-- The Nuts and Bolts

Dr. Yu (Cathy) Jiao
Applied Software Engineering Research Group
Computational Sciences and Engineering Division
Oak Ridge National Laboratory

UT-BATTELLE
Management Contractor for DOE's
Oak Ridge National Laboratory

Outline

- Introduction
- Supervised Learning vs. Unsupervised Learning
- Basic Vector Space Model (VSM)
- Similarity Measures
- Term Weighting Schemes
- Summary & Other Research Directions

Introduction

- Data Mining
  - "the nontrivial extrac
    and potentially useful infor from data"[1]
  - "the science of extracting useful information from large data sets or databases"[2]
- Data Mining Techniques
  - Supervised Learning
  - Unsupervised Learning

We are not talking about simple "keyword searching!"

Supervised Learning

Blockbuster® Entertainment mines its video rental history database to recommend rentals to individual customers.

Supervised Learning

Wal-Mart® uses massive data mining to transform its supplier relationships.

- 2,900+ stores
- 6 countries
- 3,500+ suppliers
- massive 7.5 terabyte data warehouse

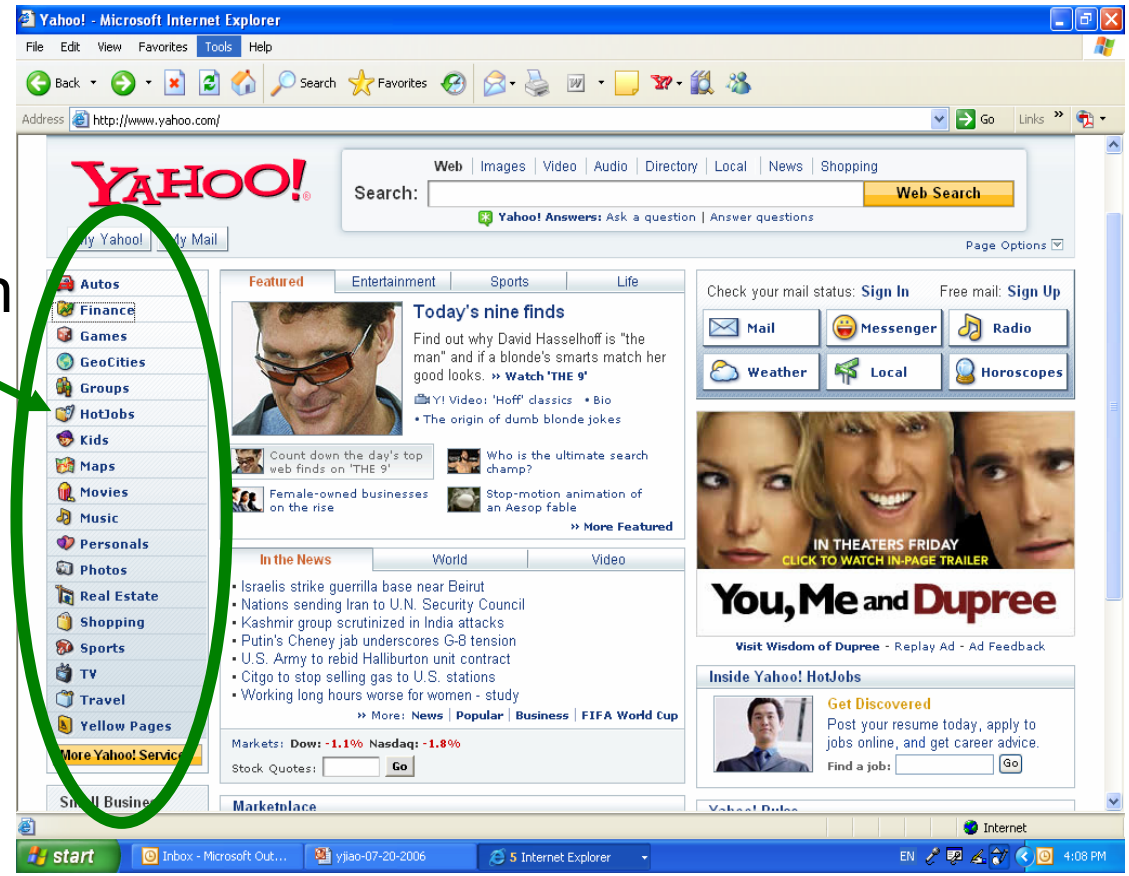
✓ Identify buying patterns

✓ Manage local store inventory

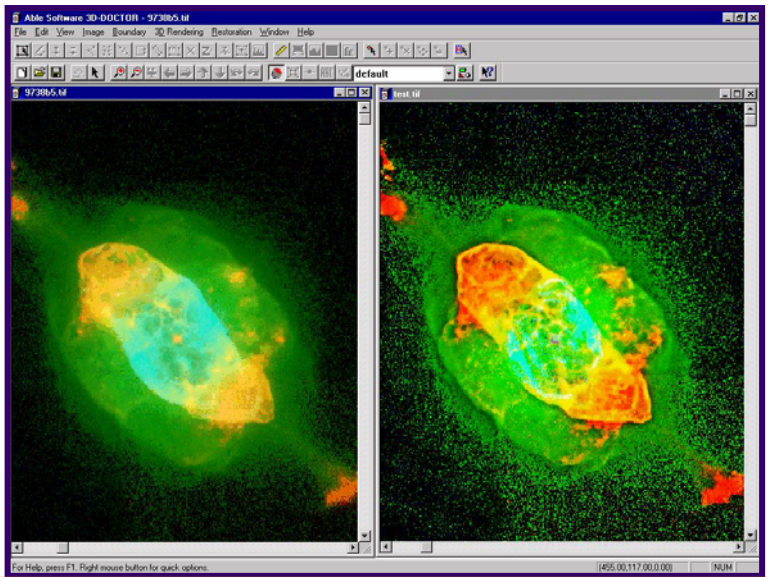
✓ Uncover new merchandising opportunities

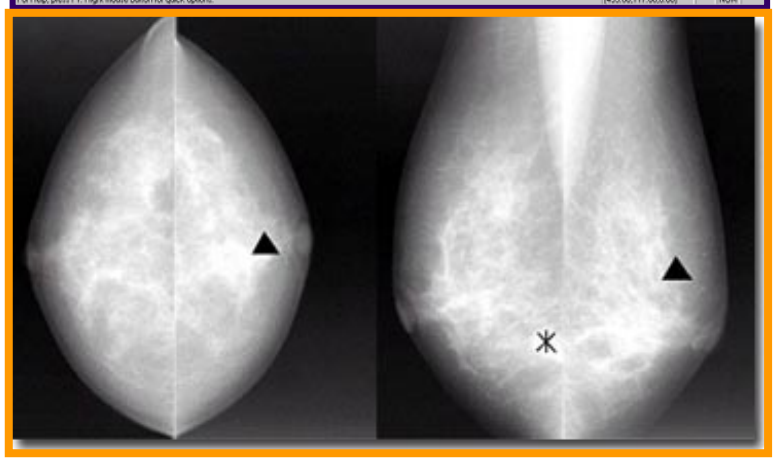
Supervised Learning

Data Classification

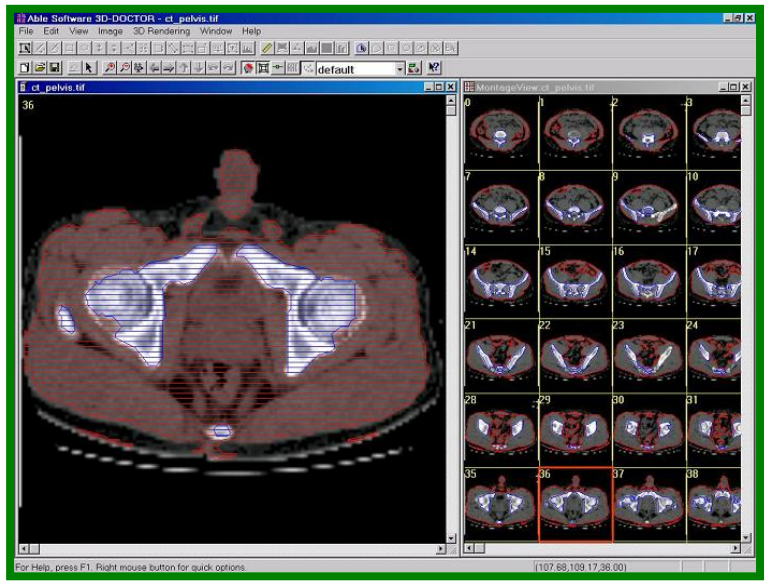


Supervised Learning



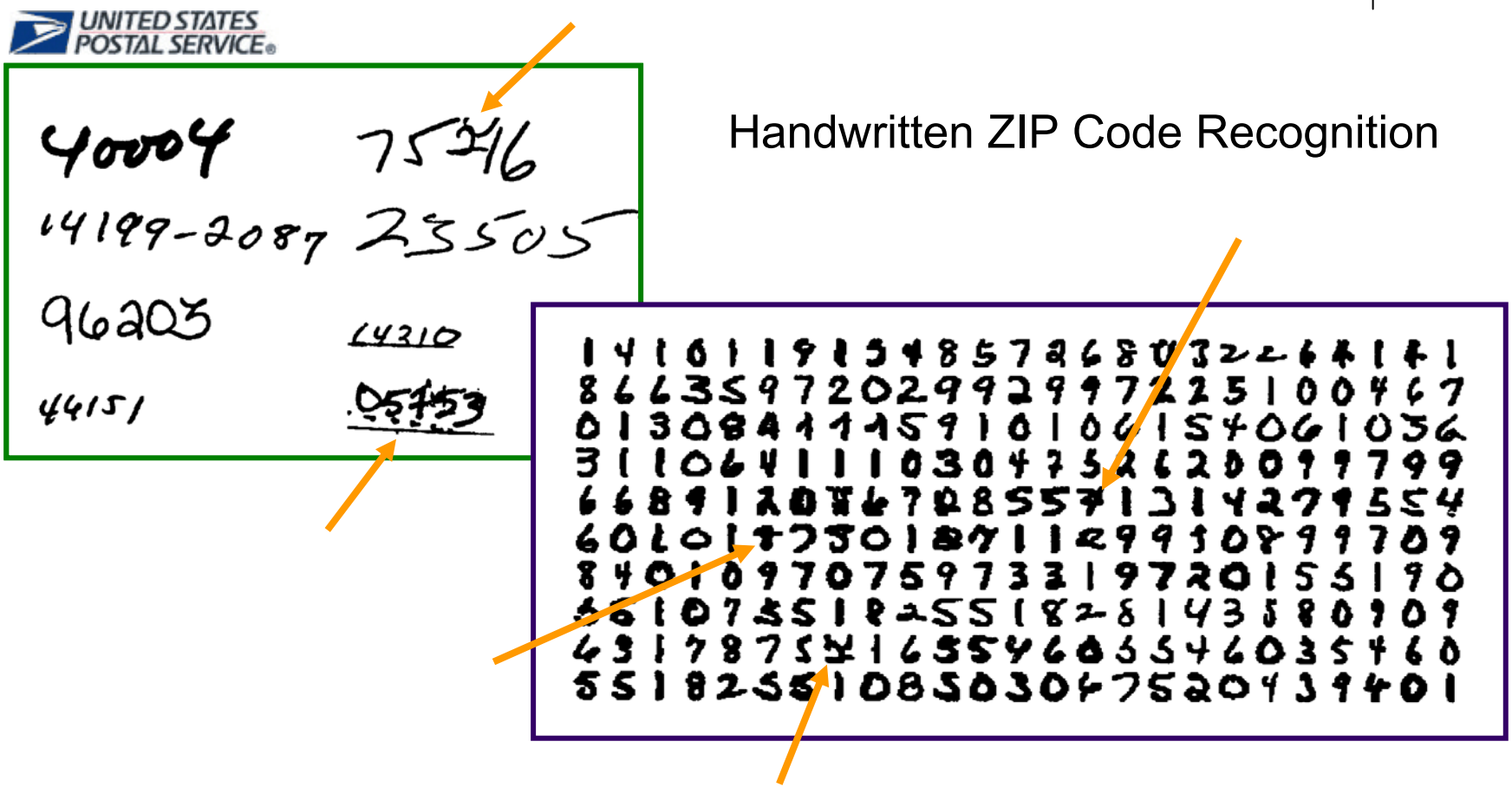


Computer-Aided Diagnosis Systems



What are they doing?

Supervised Learning

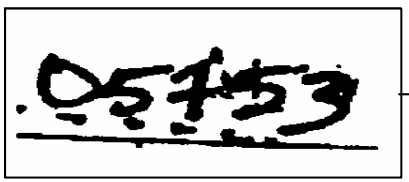


Handwritten ZIP Code Recognition

Supervised Learning

- The learning process is guided by the outcome variable. There is a measurement of the outcome.

training set → observe the outcome and feature measurements → build a prediction model (learner)

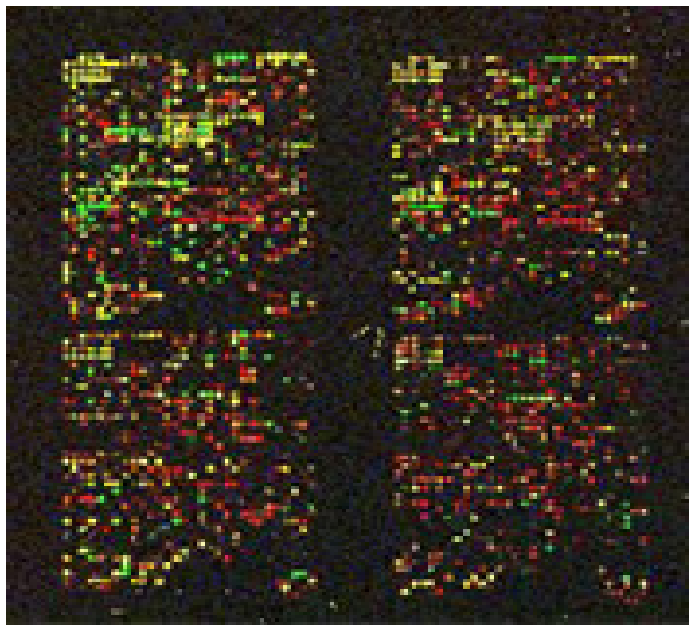
 → prediction model (learner) → 05753

Unsupervised Learning

- We observe only the features and have no measurements of the outcome. Our task is rather to describe how the data are organized or clustered.

Samples

Genes

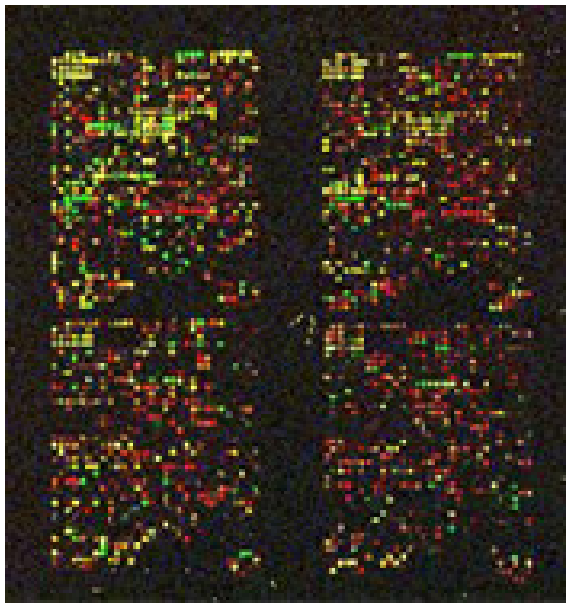


DNA micro-array data of human tumor.

Unsupervised Learning

Samples

Genes



DNA micro-array data of human tumor.

- Which samples are most similar to each other, in terms of their expression profiles across genes?
- Which genes are most similar to each other, in terms of their expression profiles across samples?
- Do certain genes show very high (or low) expression for certain cancer samples?

Unsupervised Learning

- What is in there?
- Are there any threats?
- What am I missing?

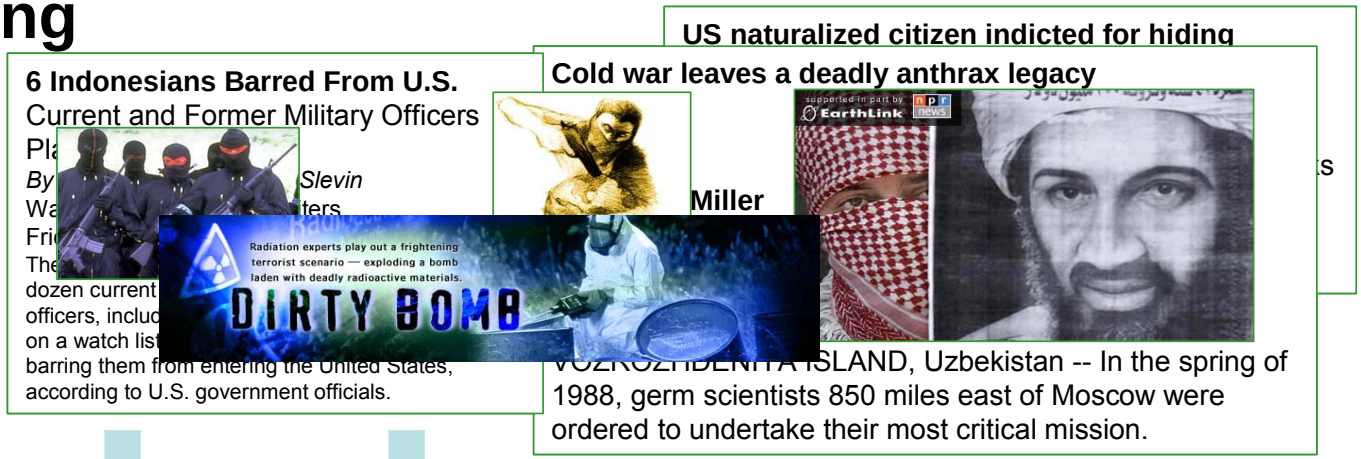
Paper Records

Old Disks

Legacy Databases

Data Clustering

Threats



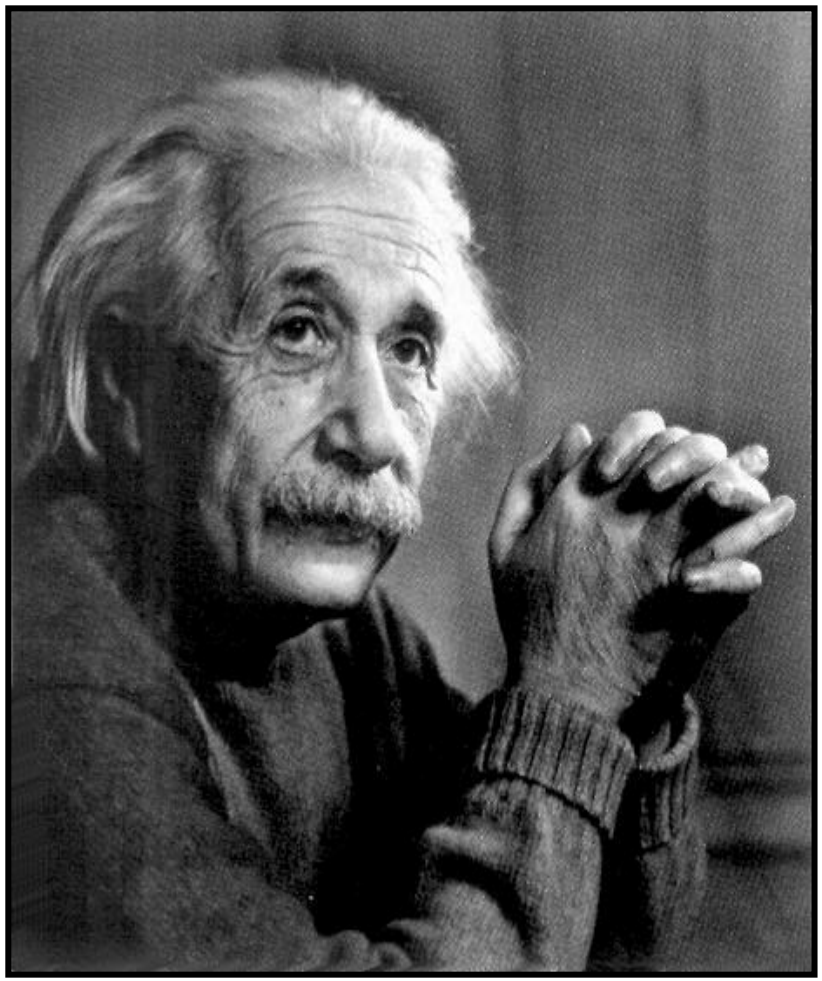
Agents On HPC

Analysis

- Swarm Intelligence
- Artificial Life
- Bayesian Statistics
- Machine Learning

Decision

Synthesizing and Disseminating Information



"If we knew what it was we were doing, it wouldn't be called research."

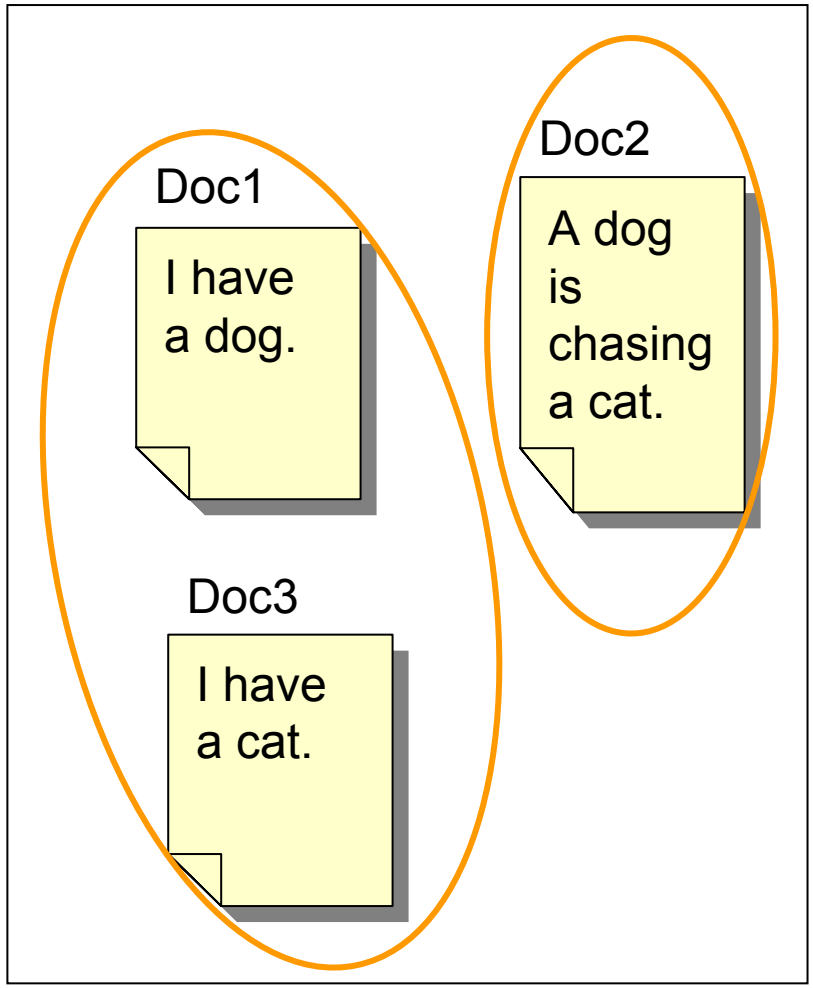
-- Albert Einstein

Summary

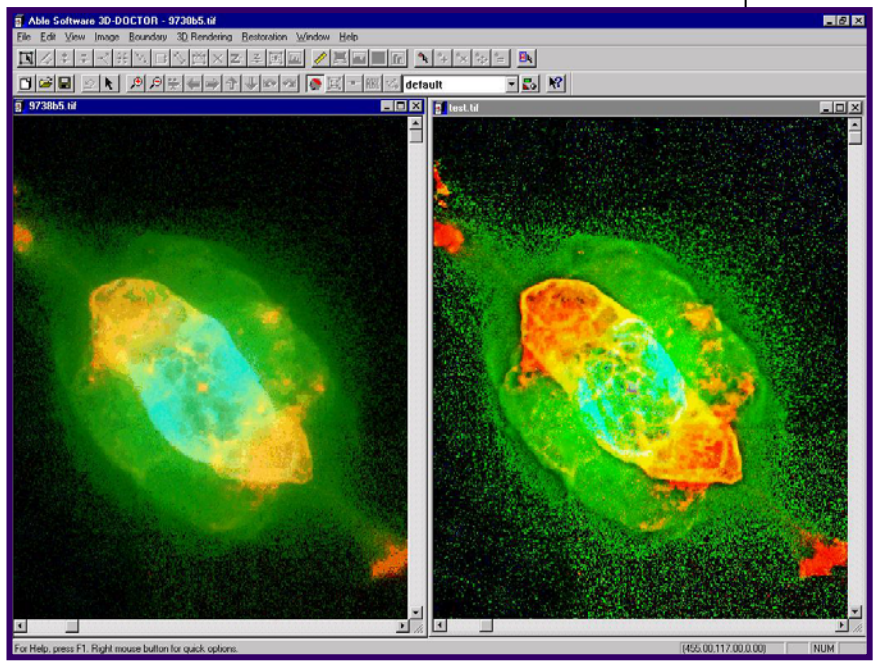
- Diverse Data Formats
  - Structured Data
  - Semi-structured Data
  - Free Text Data
  - Multimedia Data (images, videos, audio etc.)
  - …
- Various Tasks
  - Data Clustering
  - Data Classification
  - Pattern Recognition
  - Trend Analysis
  - …

Challenges

Unsupervised Learning



Supervised Learning



- How to represent data?
- How to measure similarity?

Outline

- Introduction
- Supervised Learning vs. Unsupervised Learning
- Basic Vector Space Model (VSM)
- Similarity Measures
- Term Weighting Schemes
- Summary & Other Research Directions

The Vector Space Model

- First proposed by Salton, Wong, and Yang at Cornell University in 1975 [3].
- The Vector Space Model (VSM) was originally proposed for document indexing and information retrieval.
- VSM provides a way to represent documents through the words they contain.

The Vector Space Model

- VSM defines a multi-dimensional vector space, in which each unique term of the documents is a dimension.

Doc1

I have a dog.

Doc2

A dog is chasing a cat.

Doc3

I have a cat.

Unique Terms
a
cat
chasing
dog
have
I
is

A seven-dimensional vector space.

The Vector Space Model

- Each document is represented by a seven-dimensional vector.

Doc1

I have a dog.

Doc2

A dog is chasing a cat.

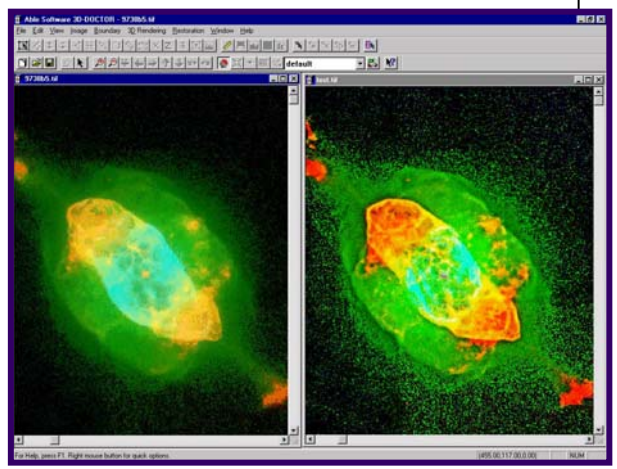
Doc3

I have a cat.

	Doc1	Doc2	Doc3
a	1	1	1
cat	0	1	1
chasing	0	1	0
dog	1	1	0
have	1	0	1
I	1	0	1
is	0	1	0

The Vector Space Model

- VSM also applies to any set of entities identified by weighted property vectors.



- The RGB color model
- Typically, computer display adapters use 8 bits for each color, giving a range of 256 possible values, or intensities, for each hue.
- We can represent each pixel in a three-dimensional space, with features red, green and blue, and use the intensities as the weights of each feature, e.g.,

  $pixel_1$[(R)124, (G) 200, (B) 30]

  $pixel_2$[(R)55, (G) 134, (B) 80]

  …

Outline

- Introduction
- Supervised Learning vs. Unsupervised Learning
- Basic Vector Space Model (VSM)
- Similarity Measures
- Term Weighting Schemes
- Summary & Other Research Directions

Similarity Measures

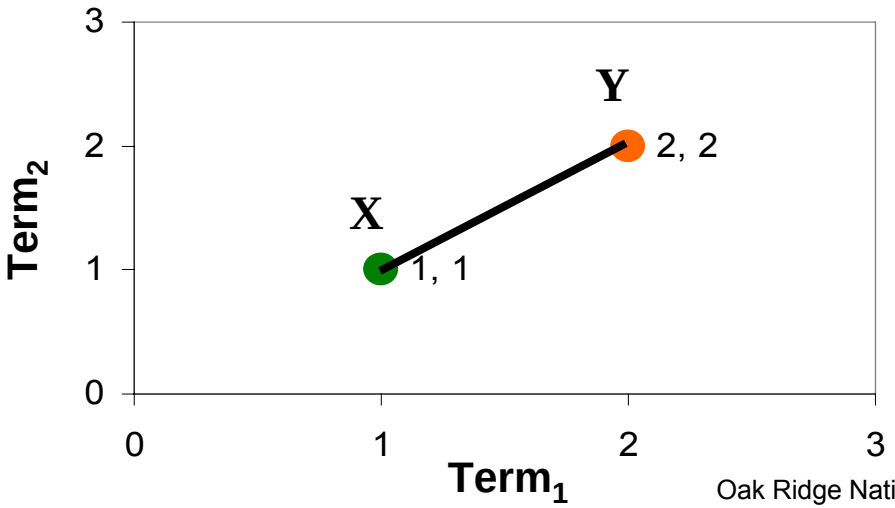
- Euclidean Distance
- Cosine Measure
- …

Euclidean Distance

- Given two vectors X(x_1, x_2, …, x_n) and Y(y_1, y_2,…, y_n),

$$EuclideanDistance(X,Y) = \sqrt{\sum (xi - yi)^2}$$

Example:



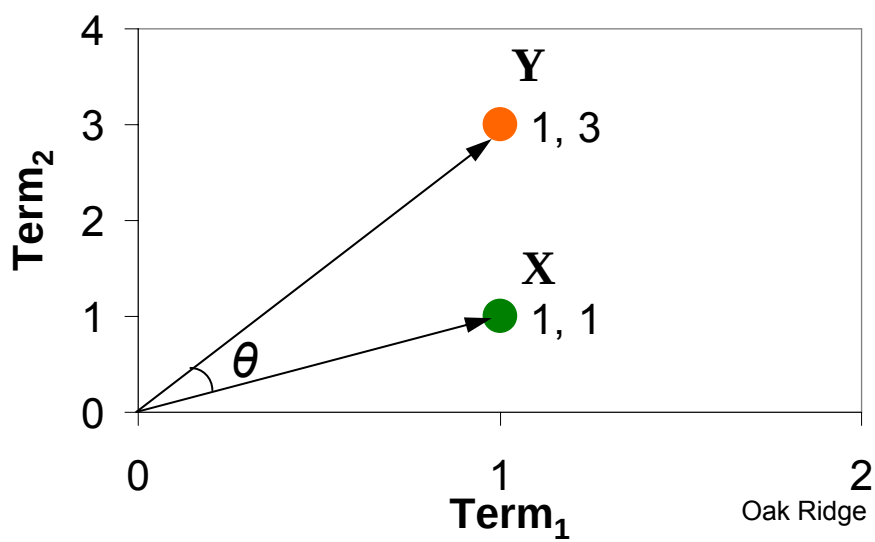
$$EuclideanDistance(X,Y)$$
$$= \sqrt{(1-2)^2 + (1-2)^2}$$
$$\approx 1.414$$

Cosine Measure

- Given two vectors X(x_1, x_2, …, x_n) and Y(y_1, y_2,…, y_n),

$$Cos\,\theta = \frac{X \bullet Y}{\| X \| \times \| Y \|} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \times \sqrt{\sum y_i^2}}$$

Example:

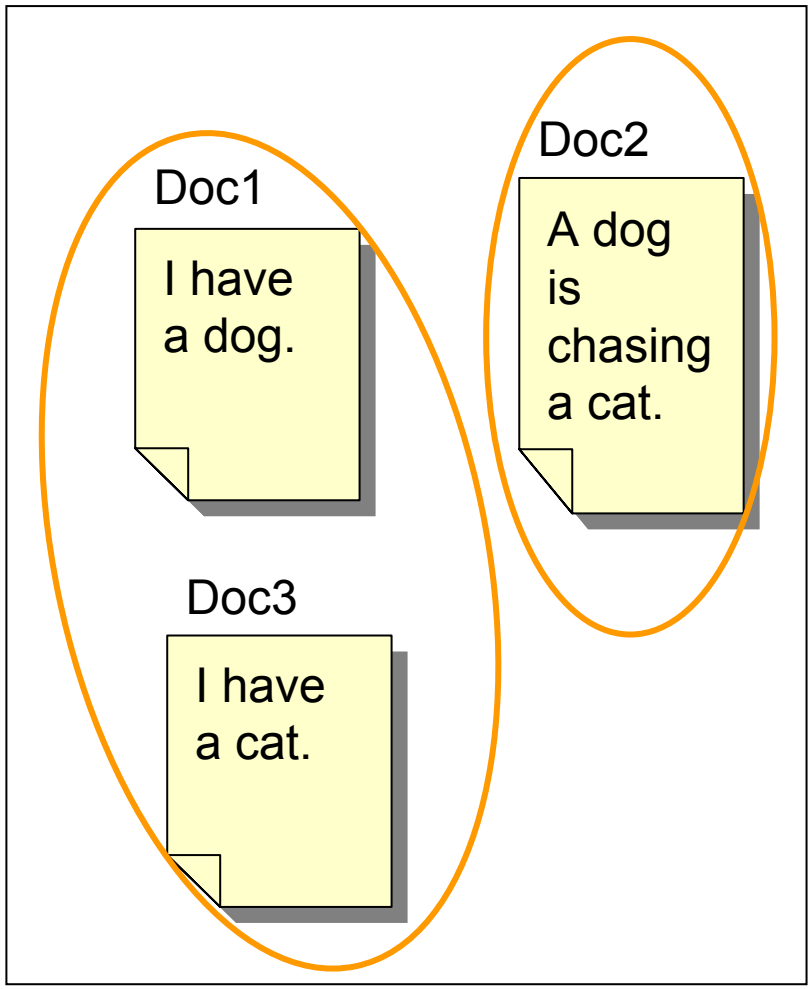


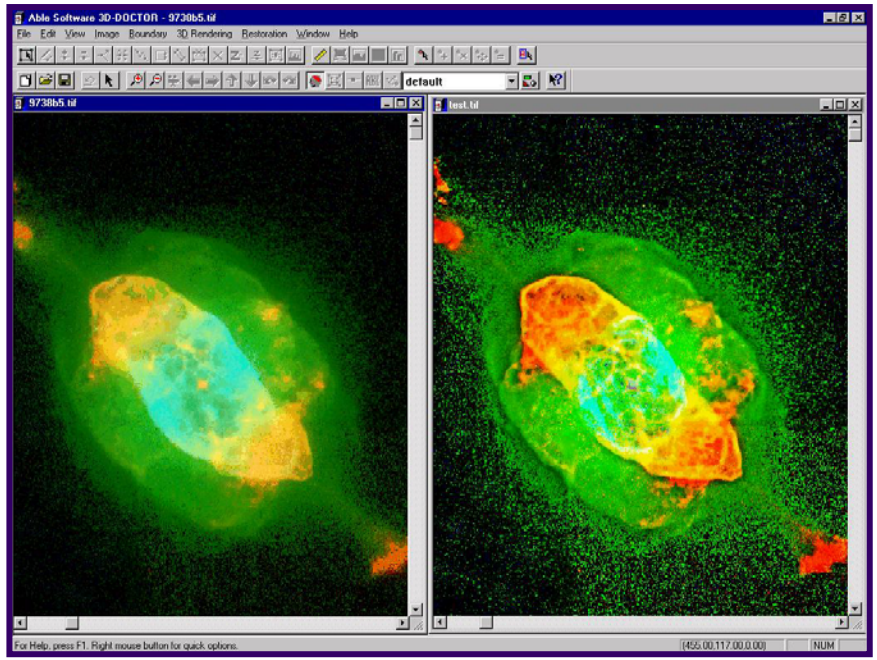
$$Cos\,\theta = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \times \sqrt{\sum y_i^2}}$$

$$= \frac{(1\times 1) + (1\times 3)}{\sqrt{1^2 + 1^2} \times \sqrt{1^2 + 3^2}}$$

$$\approx 0.89$$

Euclidean Distance vs. Cosine Measure





Doc1: “I have a dog.”

Doc2: “A dog is chasing a cat.”

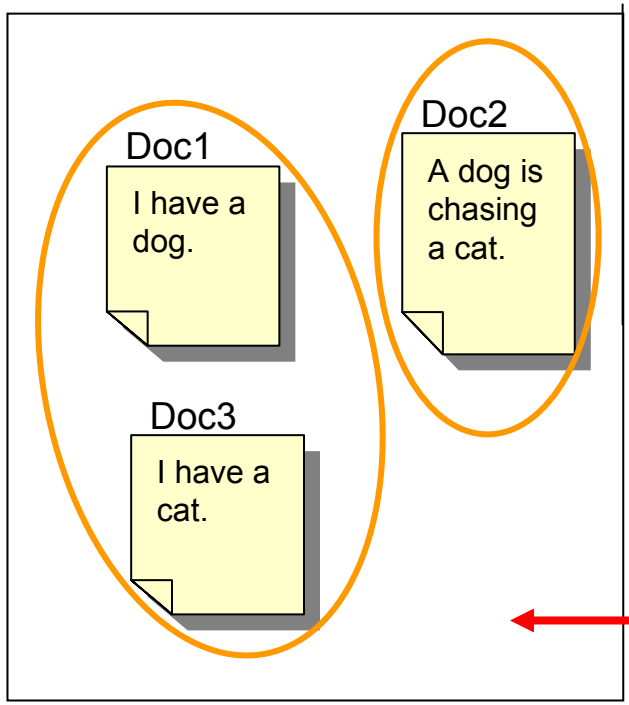
Doc3: “I have a cat.”

	Doc1	Doc2	Doc3
a	1	1	1
cat	0	1	1
chasing	0	1	0
dog	1	1	0
have	1	0	1
I	1	0	1
is	0	1	0

Doc1(1,0,0,1,1,1,0)

Doc2(1,1,1,1,0,0,1)

Doc3(1,1,0,0,1,1,0)



Cosine Similarity Matrix

	Doc1	Doc2	Doc3
Doc1	1.0	0.45	0.75
Doc2		1.0	0.45
Doc3			1.0

Outline

- Introduction
- Supervised Learning vs. Unsupervised Learning
- Basic Vector Space Model (VSM)
- Similarity Measures
- Term Weighting Schemes
- Summary & Other Research Directions

Preprocessing
-- Stopword Removal

- Stopwords are those words that appear in almost every document and have very little contribution to describing information.

Preprocessing -- Stemming

Term	Stemmed Term
uniparous	unipar
vincennes	vincenn
running	run
alderman	alderman
gentilization	gentil
spinach	spinach
runs	run
matte	matt
aleppo	aleppo
jagged	jag
metagenesis	metagenesi
body	bodi
winning	win
prelocalization	preloc
scarry	scarri

- Porter Stemmer [4]
- Lovins Stemmer [5]
- …

Term Weighting Schemes

- Terms that describe the thematic information of a document are important.
  - Term Frequency Factor (TF)
- Terms that differentiate one document from another are important.
  - Inverse Document Frequency Factor (IDF)
- Term Weighting Schemes:
  - Combine TF and IDF

Term Frequency Factor

- If a (non-stopword) term *t* appears often in a document, it is important. In other words, frequent non-stopword terms are thematic.

$f_{i,j}$: the term frequency of term *j* in document *i*.

$tf_{i,j}$: the term frequency factor of term *j* in document *i*.

$$tf_{i,j} = f_{i,j}$$

$$tf_{i,j} = \frac{f_{i,j}}{\max_j f_{i,j}}$$

$$tf_{i,j} = 0.5 + \frac{0.5 \times f_{i,j}}{\max_j f_{i,j}}$$

$$tf_{i,j} = k + \frac{(1-k) \times f_{i,j}}{\max_j f_{i,j}}$$

$$tf_{i,j} = \log(1 + f_{i,j})$$

Inverse Document Frequency (IDF) Factor

- A term's scarcity across the collection is a measure of its importance.
  - Zipf's Law: importance is inversely proportional to frequency of occurrence across documents

$$\text{DocumentFrequency} = n_j / N \approx 1/\text{rank}$$

$$\text{InverseDocumentFrequency} = N/n_j \approx \text{rank}$$

n_j: the number of documents containing *term j*

N: the total number of documents in the collection

$$idf_j = \log(1 + \frac{N}{n_j})$$

$$idf_j = \log(\frac{N - n_j}{n_j})$$

Common Term Weighting Schemes

Name	Term Weighting Scheme
TF	$w_{i,j} = tf_{i,j}$
Mutual Information[6]	$w_{i,j} = \log \dfrac{\dfrac{f_{i,j}}{N}}{\dfrac{\sum_{i=1}^{N} f_{i,j}}{N} \times \dfrac{\sum_{j=1}^{M} f_{i,j}}{N}}$
TF-IDF[3]	$w_{i,j} = tf_{i,j} \times \log(1 + N / nj)$

Common Term Weighting Schemes

Name	Term Weighting Scheme
LTU[7]	$w_{i,j} = \dfrac{(\log(tf_{i,j}) + 1.0)\log\left(\dfrac{N}{nj}\right)}{0.8 + 0.2 \times \dfrac{dl}{avg_dl}}$
Okapi[8]	$w_{i,j} = \left(\dfrac{tf_{i,j}}{0.5 + 1.5 \times \dfrac{dl}{avg_dl} + tf_{i,j}}\right)\log\left(\dfrac{N - nj + 0.5}{tf_{i,j} + 0.5}\right)$
ATC[8]	$w_{i,j} = \dfrac{\left(0.5 + 0.5 \times \dfrac{tf_{i,j}}{\max_j tf_{i,j}}\right)\log\left(\dfrac{N}{nj}\right)}{\sqrt{\sum_{i=1}^{N}\left[\left(0.5 + 0.5 \times \dfrac{tf_{i,j}}{\max_j tf_{i,j}}\right)\log\left(\dfrac{N}{nj}\right)\right]^2}}$

Limitations of TF-IDF

- It becomes inefficient when dealing with dynamic data streams.
  - Dependency on *N* (total number of documents)
  - Dependency on n_j (total number of documents containing term *j*)
  - Dependency on *avg_dl* (average document length)

Data Stream Mining

- Mining data streams is concerned with extracting knowledge structures represented in models and patterns in continuous streams of information.

- Major Challenges:
  - One-pass through data points
  - Memory constraints

Document Stream Clustering

Inverse Document Frequency

$$W_{ij} = \log_2(f_{ij}+1) * \log_2\left(\frac{N}{n}\right)$$

$$d(x_i, x_j) = \left(\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2\right)^{1/2}$$

$$O(n^2 \log n)$$

Vector Space Model

	Doc 1	Doc 2	Doc 3
Army	1	0	0
Sensor	1	1	1
Technology	1	1	0
Help	1	0	0
Find	1	0	0
Improvise	1	0	0
Explosive	1	0	1
Device	1	0	1
ORNL	0	1	0
develop	0	1	1
homeland	0	1	1
Defense	0	1	1
Mitre	0	0	1
won	0	0	1
contract	0	0	1

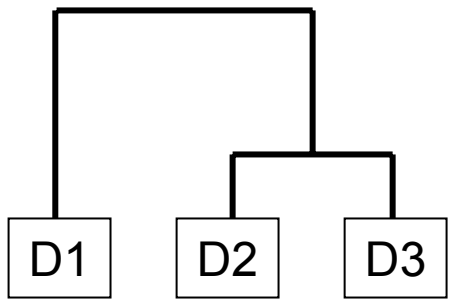
Words to Documents

Similarity Matrix

	Doc 1	Doc 2	Doc 3
Doc 1	100%	17%	21%
Doc 2		100%	36%
Doc 3			100%

Documents to Documents

Cluster Analysis



Most similar documents

Cannot work with data streams!
Need a new weighting scheme!

The TF-ICF Term Weighting

- Term Frequency – Inverse Corpus Frequency Weighting Scheme

- Objective
  - Remove the dependency on (inverse) document frequency, n_j / N .

- Questions
  - Can we use the document frequency (c_j / C) of a known corpus to approximate that of an unknown data stream?
  - Can we compose a corpus that covers most of the commonly used words?

The TF-ICF Term Weighting

- Can we use the document frequency (c_j / C) of a known corpus to approximate that of an unknown data stream?

Reference Corpora	Document Count
TREC-AP	239,055
TREC-FBIS	130,392
TREC-LATIMES	127,732
TREC-SJM	73,748
TREC-WSJ	161,576
RSS News Feeds	202,991
Total	**935,494**

Statistics:

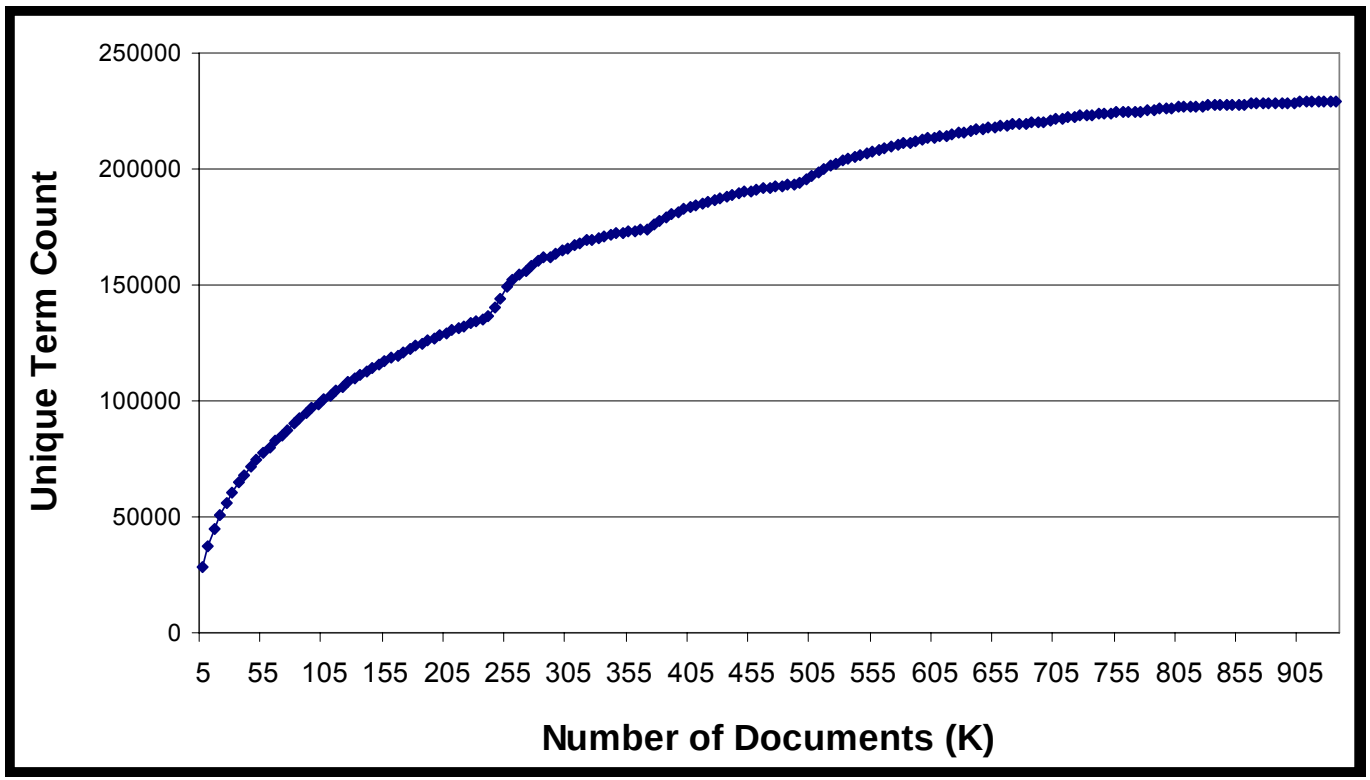
- # of Unique Terms: 229,023
- # of Common Terms: 28,872
- Average Standard Deviation of the Document Frequency (c_j / C) of the Six Corpora: < 0.01%

$$W_{ij} = \log_2(tf_{ij} + 1) * \log_2\left(\frac{C+1}{c_j+1}\right)$$

Inverse Corpus Frequency

The TF-ICF Term Weighting

- Can we compose a corpus that covers most of the commonly used words?



The TF-ICF Term Weighting

- Meet the Challenges
  - One-pass through data points
  - Memory constraints
- Moreover
  - Reduces the computational complexity of encoding *N* dynamic documents from $O(N^2)$ to $O(N)$!
  - Provides the technical foundation for parallel data stream clustering.
- Effectiveness Evaluation?

Effectiveness Evaluation

- F-Score is a statistical measure of a test's accuracy.

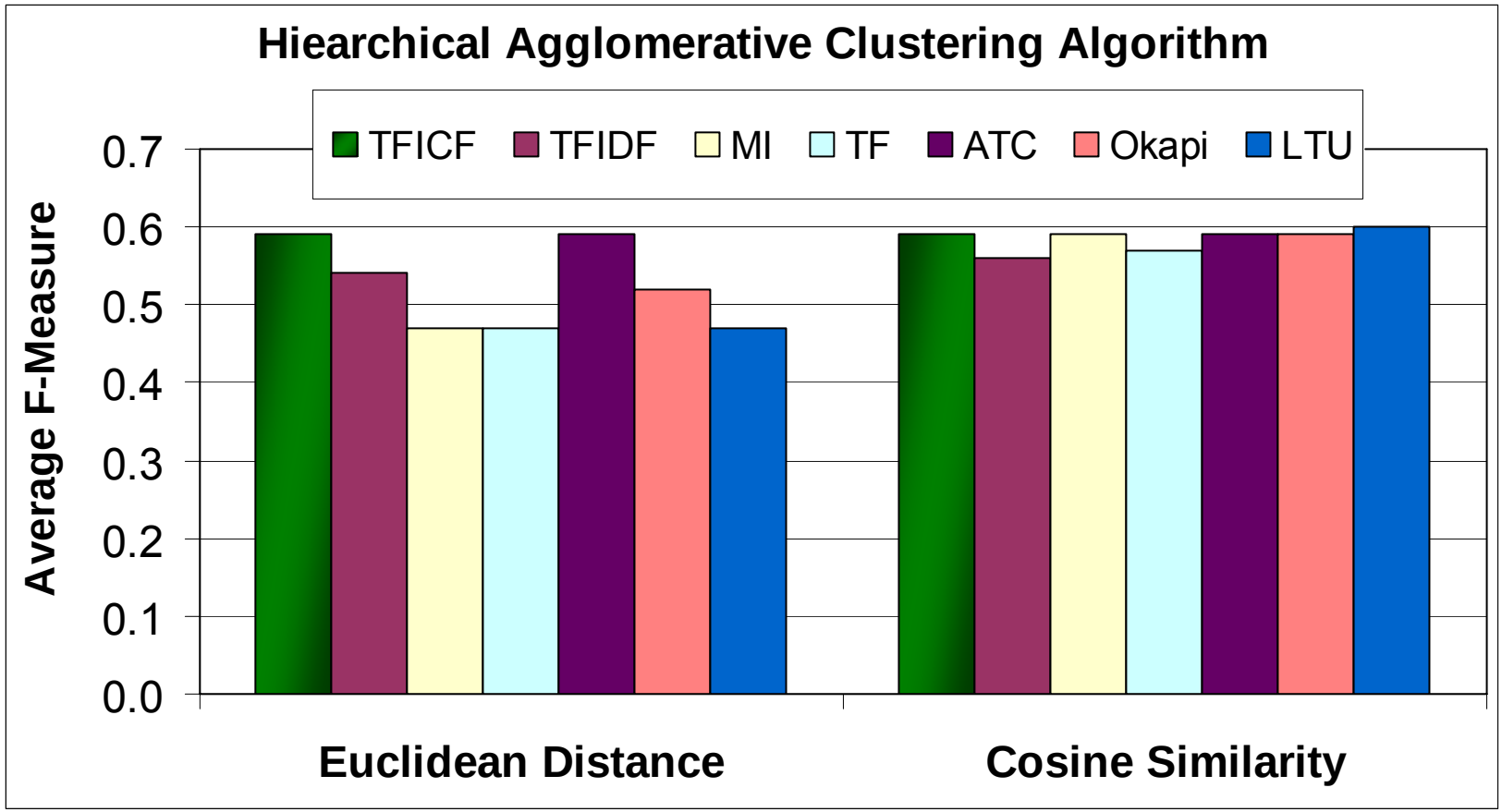
$$F-score = \frac{2pr}{p+r}$$

p: precision
r: recall

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Effectiveness Evaluation



Summary

- Supervised Learning vs. Unsupervised Learning
- VSM as a common data model for computer-understandable data representation
- Many similarity measures
- TF-IDF based term weighting schemes and their limitations
- A new term weighting scheme – TF-ICF

Other Research Directions

- Swarm Intelligence & Data Clustering
  – Dr. Xiaohui Cui (July 27th)
- Parallel Data Mining Algorithms
- Efficient Dimensionality Reduction
- Information Summarization
- Information Fusion
- Information Visualization
- …

References

1. W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview. *AI Magazine*, Fall 1992, pp. 213-228.
2. D. Hand, H. Mannila, P. Smyth: Principles of Data Mining. MIT Press, Cambridge, MA, 2001. ISBN 0-262-08290-X
3. G. Salton, A. Wong, and C.S. Yang, A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 1975, 18(11) pp. 613-620.
4. M.F. Porter, An Algorithm for Suffix Stripping, *Program*, 1980, 14(3), pp. 130-137.
5. J.B. Lovins, Development of a Stemming Algorithm, *Mechanical Transformation and Computational Linguistics*, 1968, 11, pp. 22-31.
6. Pantel, P. and Lin D. Document clustering with committees. In *Proc. of the 25th ACM International Conference on Research and Development in Information Retrieval* (*SIGIR'02*), 199-206.
7. Buckley, C., Singhal, A., and Mitra, M. New retrieval approaches using SMART. In *Proc. of the 4th Text Retrieval conference (TREC-4)*, Gaithersburg, 1996.
8. Jin, R., Falusos, C., and Hauptmann, A.G. Meta-scoring: automatically evaluating term weighting schems in IR without precision-recall. In *Proc. of the 24th ACM International Conference on Research and Development in Information Retrieval* (*SIGIR'01*), 83-89, 2001.
9. D'Anna, C. A., Zechmeister, E. B., and Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. Journal of Reading Behavior, 23(1), 109-122.
10. Goulden, R., Nation, P., and Read, J. (1990). How large can a receptive vocabulary be? Applied Linguistics, 11(4), 341-363.
11. Hazenberg, S., and Hulstijn, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native students: An empirical investigation. Applied Linguistics, 17(2), 145-163. cited in: Biemiller,, Andrew. (Spring 2001). Teaching vocabulary. American Educator. online.

Q&A

Yu (Cathy) Jiao

jiaoy@ornl.gov

More Similarity Measures?

- Jaccard Coefficient
- Pearson Correlation Coefficient
- …

Jaccard Coefficient

- The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets.

- The Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A, B) = |A \cap B| / |A \cup B|$$

Pearson's Correlation Coefficient

- In statistics, the Pearson product-moment correlation coefficient (first introduced by Francis Galton) , r, is a measure of how well a linear equation describes the relation between two variables X and Y.
- It is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom:

$$r = \frac{\sum z_x z_y}{N-1} \qquad z = \frac{X-\mu}{\sigma}.$$

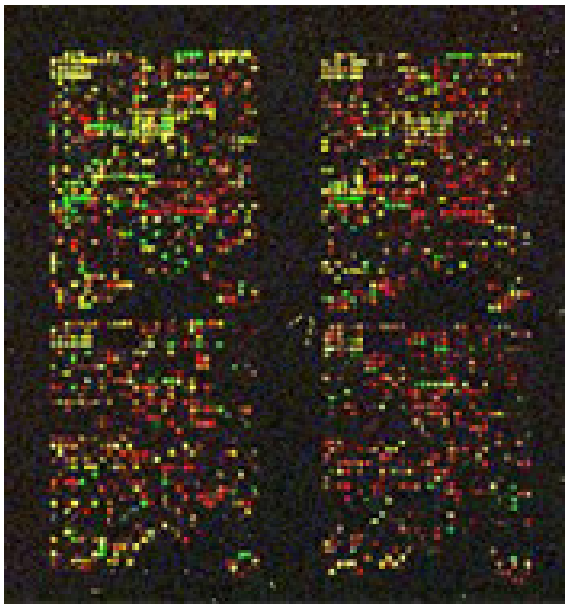
X is a raw score to be standardized
σ is the standard deviation of the population
μ is the mean of the population
N is the number of samples

- The coefficient ranges from -1 to 1.
- A value of 1 shows that a linear equation describes the relationship perfectly and positively.
- A value of -1 shows that the relationship is linear and negative.
- A value of 0 shows that a linear model is inappropriate.

DNA Micro-array

Samples

Genes



DNA micro-array data of human tumor.

- A DNA micro-array measures the expression of a gene in a cell by measuring the amount of messenger RNA present for that gene.
  - The nucleotide sequences for a few thousand genes are printed on a glass slide.
  - A target sample and a reference sample are labeled with red and green dyes, and each are hybridized with the DNA on the slide.
  - Through fluoroscopy, the log(red/green) intensities of RNA hybridizing at each site is measured. Positive values (red) indicating higher expression in the target vs. the reference, and vice versa for negative values (green).

Data, Information, and Knowledge

- Data:
  1. Information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful
  2. …
- Information
  1. Knowledge obtained from investigation, study, or instruction
  2. …
- Knowledge
  1. The fact or condition of knowing something with familiarity gained through experience or association.
  2. …

The RGB Color Model

- The RGB color model is an additive model in which red, green and blue are combined in various ways to produce other colors.
- Typically, computer display adapters use up to 24 bits of information for each pixel (8 bits for each color), giving a range of 256 possible values, or intensities, for each hue.
- Using VSM, we can represent each data point (pixel) in a three-dimensional space, with features red, green and blue, and use the intensities as the weights of each feature, e.g.,
- pixel1[(R)124, (G) 200, (B) 30].