Dimensionality Reduction

Dr Muhammad Atif Tahir

NUCES - FAST

Review Lecture 8

- Association Rules Mining
- Apriori Principle
- FP Tree

Dimensionality Reduction

Dimensionality Reduction

- The input space of many learning problems is of high dimensionality

- This has computational implications and, it makes finding the intrinsic information content difficult

- Dimensionality reduction methods usually try to address these two problems at the same time

- Context:
  - unsupervised learning
  - given a collection of data points in an n-dimensional space
  - a good representation of the data in r-dimensional

Dimensionality Reduction - Principle

- m examples; n dimensional space

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$$

$$\mathbf{X} = \begin{matrix} [\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1m}] \\ [\mathbf{x}_{21}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1m}] \\ [\ldots\ldots\ldots\ldots\ldots\ldots\ldots] \\ [\mathbf{x}_{n1}, \mathbf{x}_{n2}, \ldots, \mathbf{x}_{nm}] \end{matrix}$$

Columns are the examples.

This can always be presented by factoring the data matrix as a product of r basis vectors and m code words

When r=n, the code words are identical to the original data points

Dimensionality Reduction - Principle

- m examples; n dimensional space $\quad \mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$

$$\mathbf{X} = \begin{matrix} [\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1m}] \\ [\mathbf{x}_{21}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1m}] \\ [\ldots\ldots\ldots\ldots\ldots\ldots] \\ [\mathbf{x}_{n1}, \mathbf{x}_{n2}, \ldots, \mathbf{x}_{nm}] \end{matrix} = \begin{matrix} [1\,0\,0\ldots.0] \\ [0\,1\,0\ldots.0] \\ [0\,0\,1\ldots.0] \\ [0\;\,0\;\,0\ldots1] \end{matrix} \mathbf{x} \begin{matrix} [\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1m}] \\ [\mathbf{x}_{21}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1m}] \\ [\ldots\ldots\ldots\ldots\ldots\ldots] \\ [\mathbf{x}_{n1}, \mathbf{x}_{n2}, \ldots, \mathbf{x}_{nm}] \end{matrix}$$

$$\mathbf{n\,x\,m} \quad = \quad \mathbf{n\,x\,r} \quad \mathbf{x} \quad \mathbf{r\,x\,m}$$

$$\mathbf{X} = \mathbf{WV}$$

Columns are the examples.
This can always be presented by factoring
the data matrix as a product of r basis vectors and m code words
When r=n, the <u>code words are identical to the original data points</u>

Data Encoding

- m examples; n dimensional space

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$$

$$\mathbf{X} = \begin{matrix} [x_{11}, x_{12}, \ldots, x_{1m}] \\ [x_{21}, x_{12}, \ldots, x_{1m}] \\ [\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots] \\ [x_{n1}, x_{n2}, \ldots, x_{nm}] \end{matrix} = \begin{matrix} [\ldots w_1^T \ldots] \\ [\ldots w_2^T \ldots] \\ [\ldots\ldots\ldots\ldots] \\ [\ldots w_r^T \ldots] \end{matrix} \mathbf{x} \begin{matrix} [v_{11}, v_{12}, \ldots, v_{1m}] \\ [v_{21}, v_{12}, \ldots, v_{1m}] \\ [\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots] \\ [v_{r1}, v_{r2}, \ldots, v_{rm}] \end{matrix}$$

$$n \times m \quad = \quad n \times r \quad \mathbf{x} \quad r \times m$$

$$\min_{V,W} \|X - WV\|^2$$

Data Encoding

$$\min_{V,W} \|X - WV\|^2$$

- Given a new basis W, each data point is represented as a linear combination of the basis vectors and the goal is to minimize the reconstruction error

- Methods differ in the constraints they impose on V (the encoding matrix). Consequently, they differ computationally

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} w_{11} \cdots\cdots w_{1n} \\ w_{21} \cdots\cdots w_{2n} \\ \cdots \\ w_{n1} \cdots\cdots w_{nn} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \cdots \\ v_n \end{pmatrix}$$

Dimensionality Reduction Techniques

Visualize, categorize, or simplify large datasets.

- Principal Component Analysis: Finds the dimensions that capture the most variance

- MDS: Finds data points in lower dimensional space that best preserves the inter-point distance.

- Isomap: Estimates the distance between two points on a manifold by following a chain of points with shorter distances between them. (More accurate in representing global distances than LLE; slower than LLE)

- LLE (Local Linear Embedding): Only worries about representing the distances between local points. Faster than Isomap (no worry about global distances).
- Hesian Eigenmaps:
- Laplacian Eigenmaps
- Charting
- SOM (Kohonen's Self-organizing map)

Feature Selection

Feature Selection

- In many applications, we often encounter a very large number of potential features that can be used
- Which subset of features should be used for the best classification?
- Need for a small number of discriminative features
  - To avid “curse of dimensionality”
  - To reduce feature measurement cost
  - To reduce computational burden
- Given an nxd pattern matrix (n patterns in d-dimensional feature space), generate an nxm pattern matrix, where m << d

Feature Selection vs. Extraction

- Both are collectively known as dimensionality reduction
- Selection: choose a best subset of size m from the available d features
- Extraction: given d features (set *Y*), extract m new features (set *X*) by linear or non-linear combination of all the d features
  - Linear feature extraction: $X = TY$, where T is a mxd matrix
  - Non-linear feature extraction: $X = f(Y)$
- New features by extraction may not have physical interpretation/meaning
- Examples of linear feature extraction
  - Unsupervised: PCA; Supervised: LDA/MDA
- Criteria for selection/extraction: either improve or maintain the classification accuracy, simplify classifier complexity

Feature Selection

- How to find the best subset of size m?
- Recall, best means classifier based on these m features has the lowest probability of error of all such classifiers
- Simplest approach is to do an exhaustive search; computationally prohibitive
  - For d=24 and m=12, there are about 2.7 million possible feature subsets! Cover & Van Campenhout (IEEE SMC, 1977) showed that to guarantee the best subset of size m from the available set of size d, one must examine all possible subsets of size m
- Heuristics have been used to avoid exhaustive search
- How to evaluate the subsets?
  - Error rate; but then which classifier should be used?
  - Distance measure; Mahalanobis, divergence,…
- Feature selection is an optimization problem

Feature Selection: Evaluation, Application, and Small Sample Performance (Jain & Zongker, IEEE Trans. PAMI, Feb 1997)

- Value of feature selection in combining features from different data models
- Potential difficulties feature selection faces in small sample size situation
- Let Y be the original set of features and X is the selected subset
- Feature selection criterion function for the set X is J(X); large values of J indicates better feature subset; problem is to find subset X such that

$$J(X) = \max_{Z \subseteq Y, |Z| = d} J(Z)$$

Taxonomy of Feature Selection Algorithms

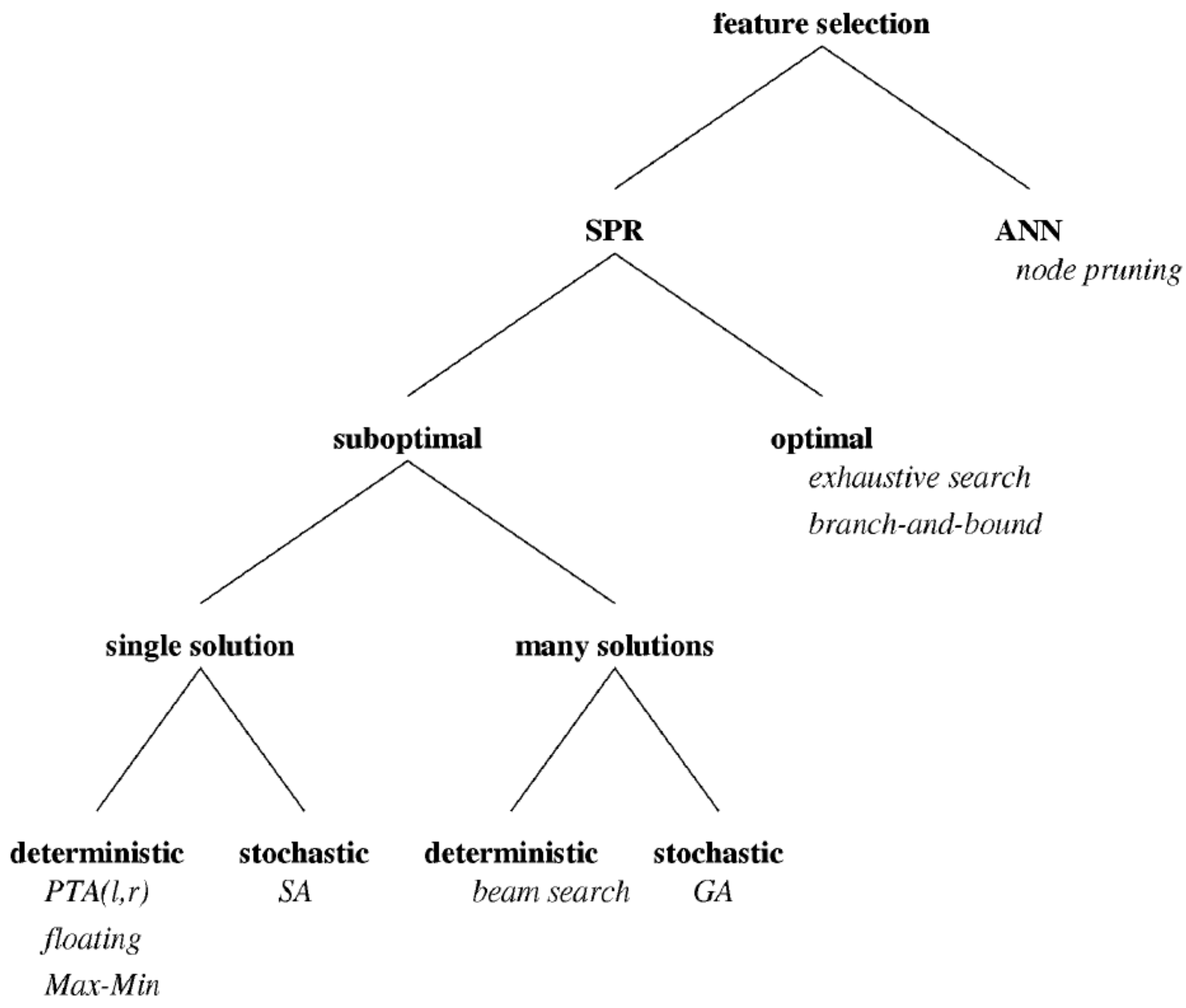


Fig. 1. A taxonomy of feature selection algorithms.

Deterministic Single-Solution Methods

- Begin with a single solution (feature subset) & iteratively add or remove features until some termination criterion is met
- Also known as sequential methods; most popular
  - Bottom up/forward methods: begin with an empty set & add features
  - Top-down/backward methods: begin with a full set & delete features
- Since they do not examine all possible subsets, no guarantee of finding the optimal subset
- Pudil introduced two floating selection methods: SFFS, SFBS
- 15 feature selection methods listed in Table 1 were evaluated

TABLE 1
FEATURE SELECTION ALGORITHMS USED IN EXPERIMENTAL EVALUATION

SFS	SBS	GSFS(2)	GSBS(2)
GSFS(3)	GSBS(3)	SFFS	SFBS
PTA((1), (2))	PTA((1), (3))	PTA((2), (3)	
BB	MM	GA	NP

Sequential Forward Selection (SFS)

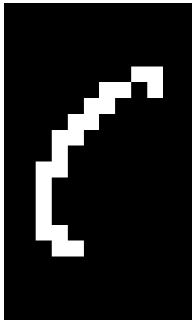
- Start with empty set, *X=0*
- Repeatedly add most significant feature with respect to *X*
- *Disadvantage: Once a feature is retained, it cannot be discarded; nesting problem*

Sequential Backward Selection (SBS)

- Start with full set, $X=Y$
- Repeatedly delete least significant feature in X
- *Disadvantage: SBS requires more computation than SFS; Nesting problem*

Multiclass, Structured and Multi-label Classification

Basic Classification in ML

	Input	Output
	$\mathbf{x} \in \mathcal{X}$	$y \in \mathcal{Y}$
Spam filtering		**Binary** 
Character recognition		**Multi-Class** **C**

Structured Classification

Input

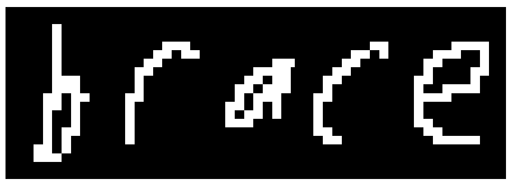
$\mathbf{x} \in \mathcal{X}$

Output

$\mathbf{y} \in \mathcal{Y}$

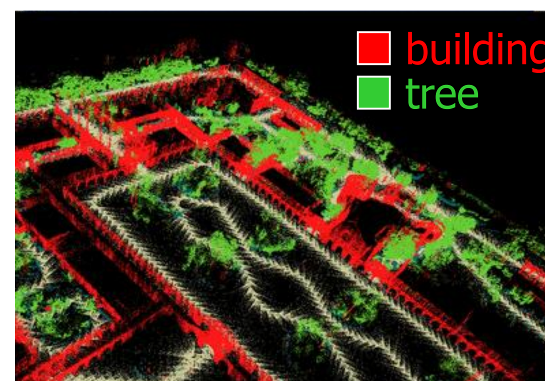
Handwriting recognition

Structured output



brace

3D object recognition



Multi-label Classification

Input	Output
$\mathbf{x} \in \mathcal{X}$	$\mathbf{y} \in \mathcal{Y}$

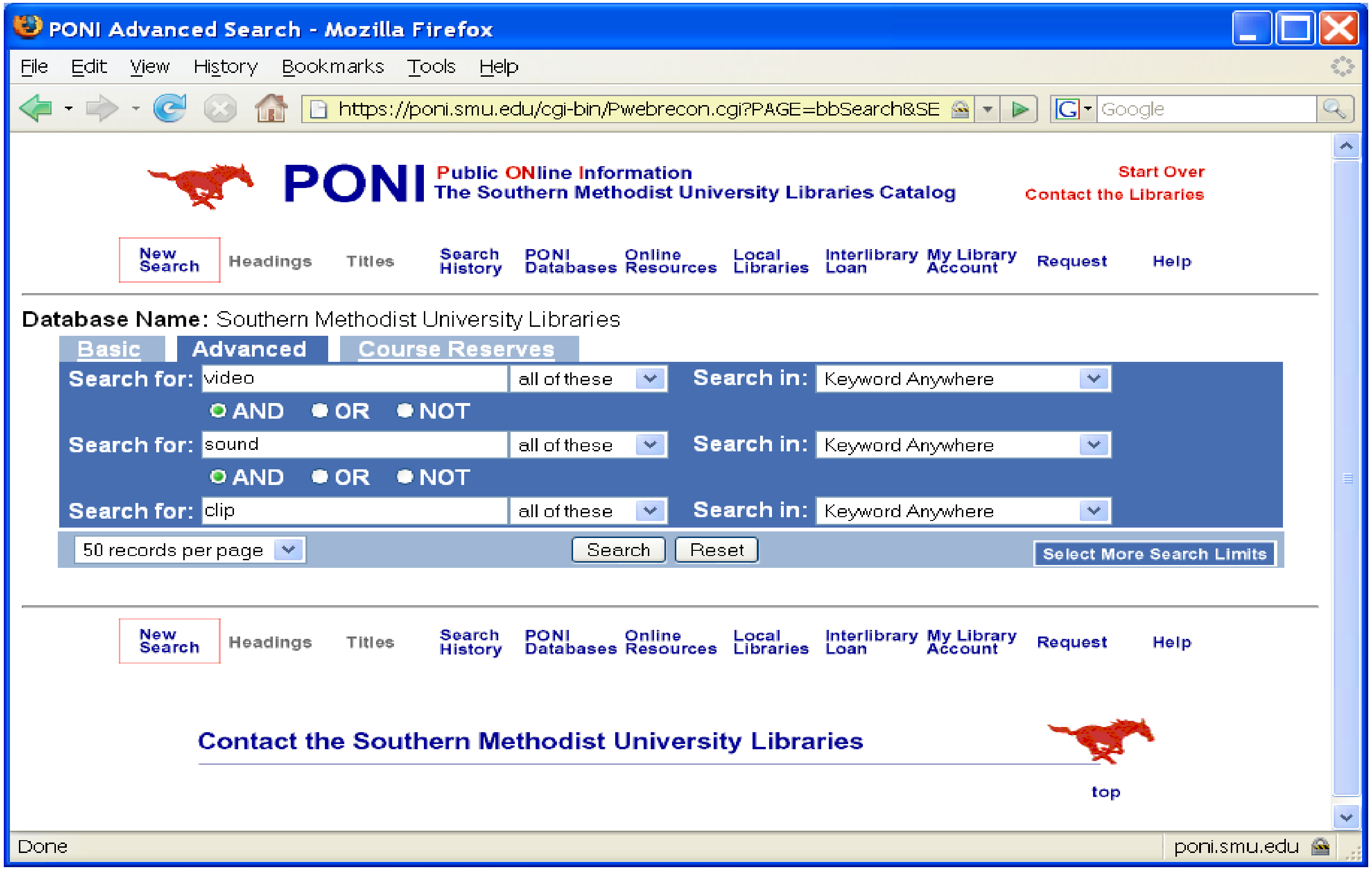
Horse, Person, Tree

Text Mining

(From Book, Chapter 15)

Image / Video Retrieval

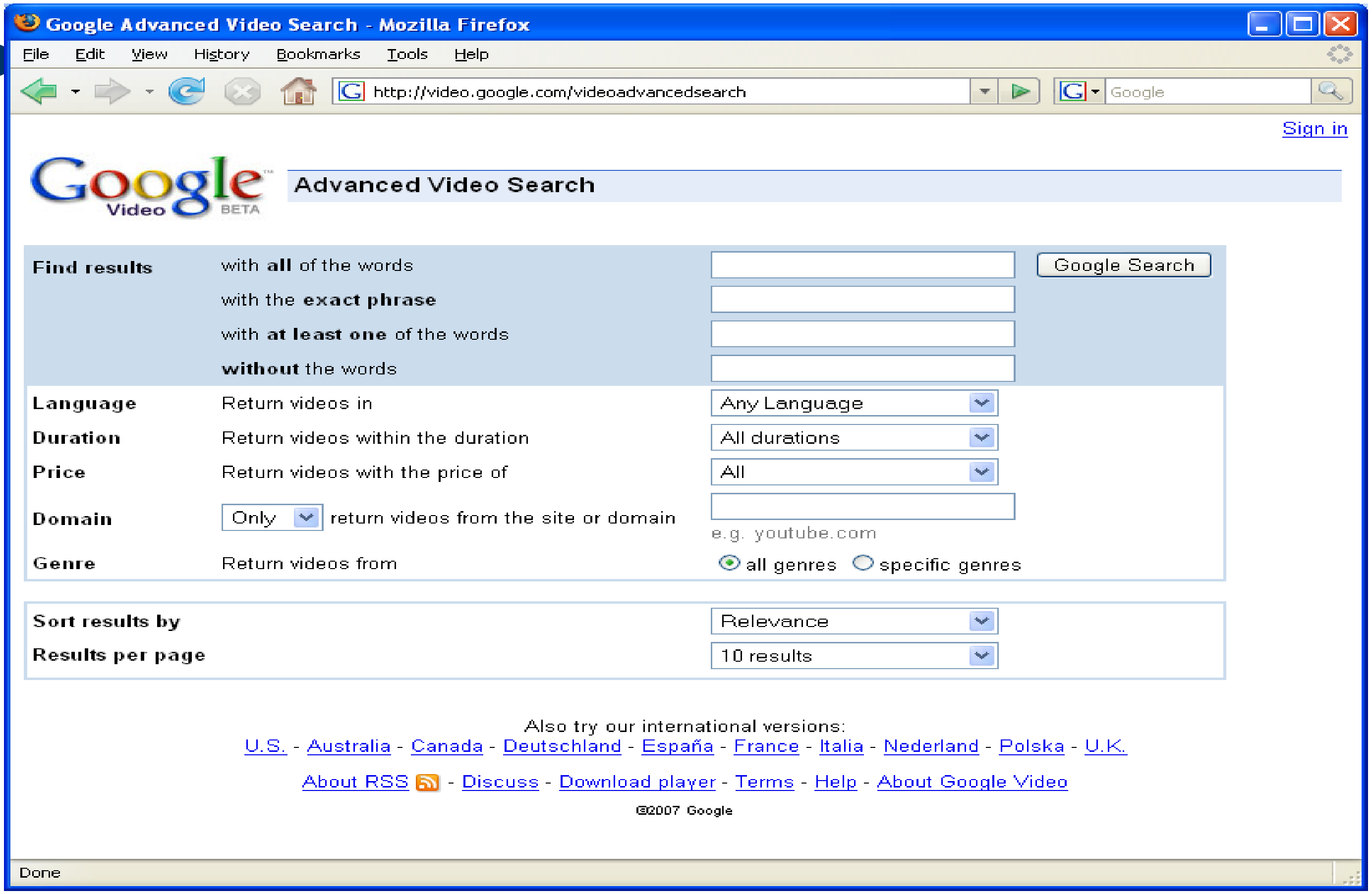
Introduction



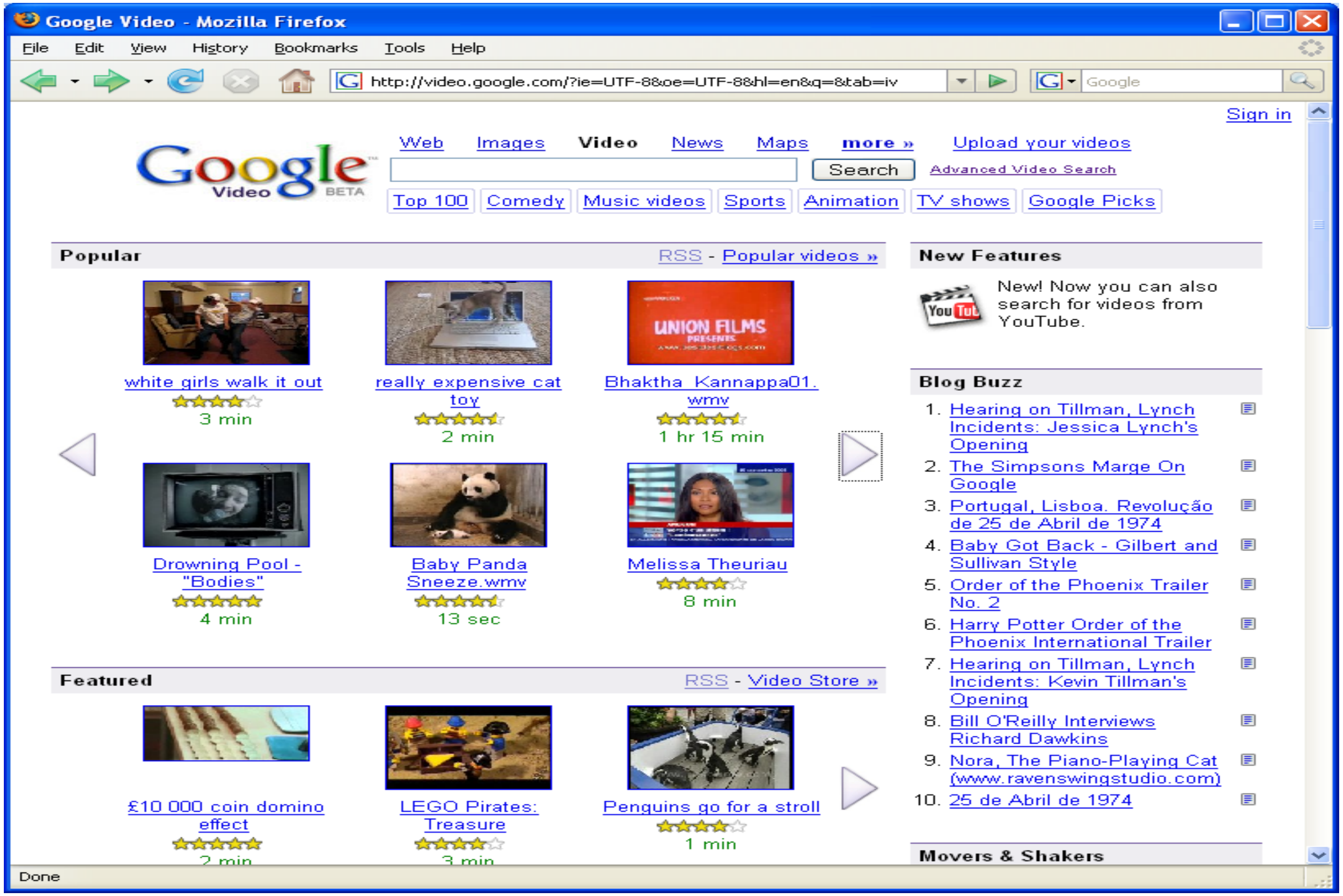
Introduction (cont.)

- Other search engines still using description search method.
- Current image search method: by description.

Introduction (cont.)



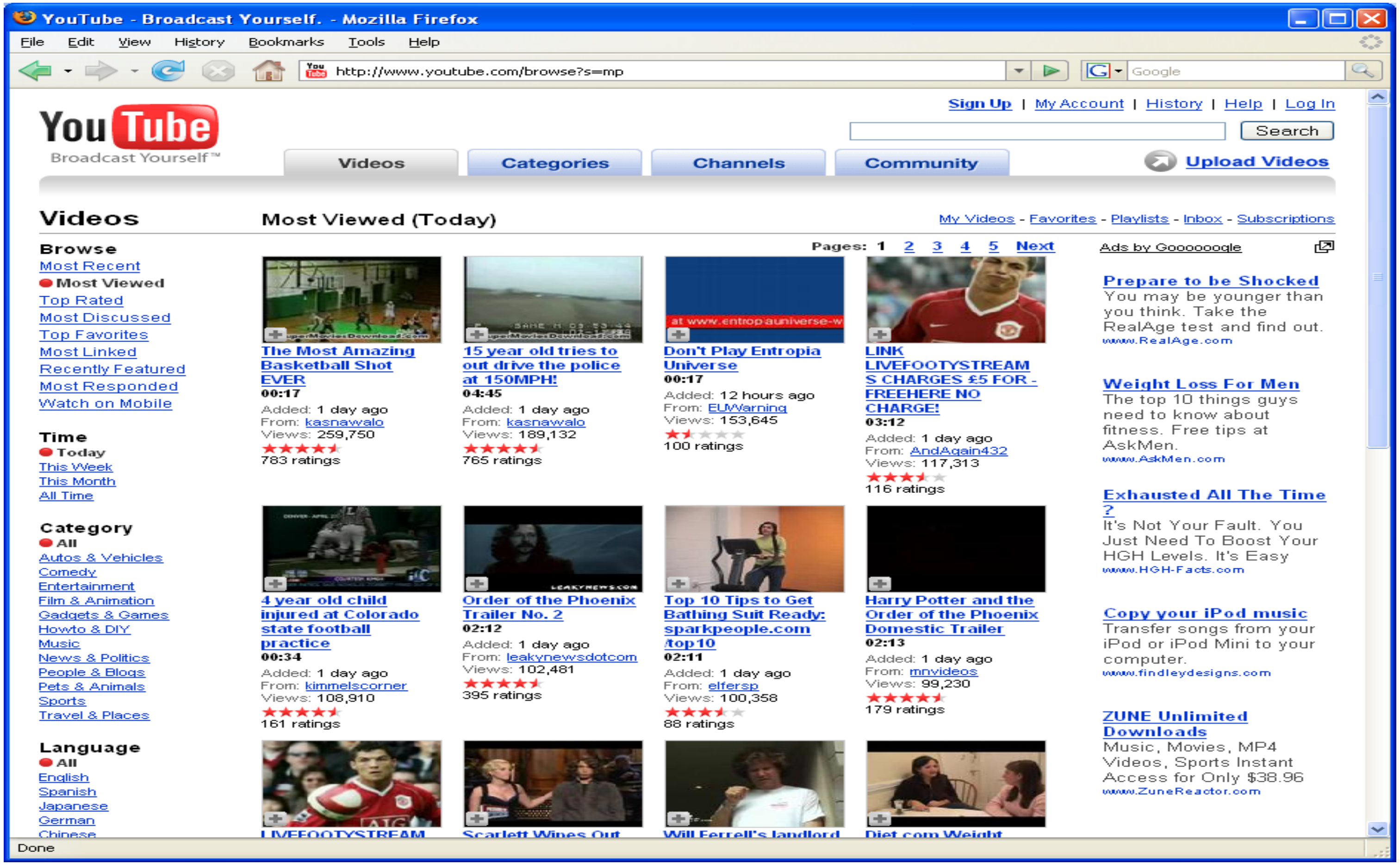
Introduction (cont.)



Introduction (cont.)

- Picture is worth a thousand words.
- More than words can express.
- Growing number video clips on MySpace and YouTube, there is a need for a video search engine.

Introduction (cont.)



Introduction (cont.)

- Therefore, we need a better search technique – Content-Based Video Retrieval System (CBVR).

Introduction (cont.)

- What good is video retrieval?
  - Historical Achieve
  - Forensic documents
  - Fingerprint & DNA matching
  - Security usage

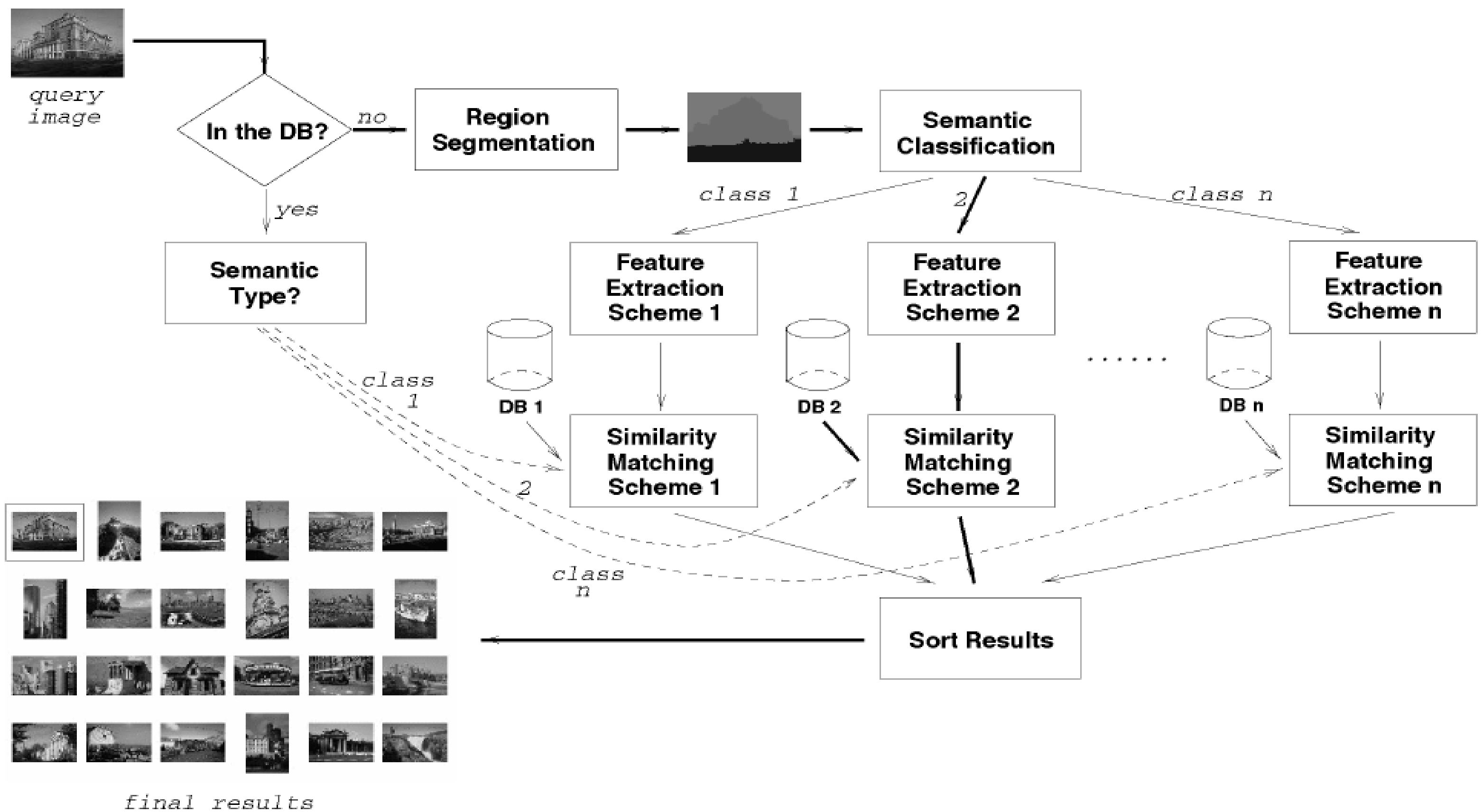
Overview (cont.)

- CBVR has two Approaches:
  - Attribute based
  - Object based
- CBVR can be done by:
  - Color
  - Texture
  - Shape
  - Spatial relationship
  - Semantic primitives
  - Browsing
  - Objective Attribute
  - Subjective Attribute
  - Motion
  - Text & domain concepts

Overview (cont.)

- CBVR has two phases:
  - Database Population phase
    - Video shot boundary detection
    - Key Frames selection
    - Feature extraction
  - Video Retrieval phase
    - Similarity measure

Overview (cont.)



[Wang, Li, Wiederhold, 2001]

References

- Professor Dan Roth Lecture Notes on Dimensionality Reduction, University of Illinois at Urbana Champaign
- Content-Based Video Retrieval System, Lecture Notes by Edmund Liang
- Lecture Notes By Anil Jain (MSU) on Feature Selection

Questions!