Multimedia Data

- Description-based retrieval systems
  - Build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation
- Content-based retrieval systems
  - Support retrieval based on the image content, such as color histogram, texture, shape, objects, and wavelet transforms

Queries in Content-Based Retrieval Systems

- Image sample-based queries
  - Find all of the images that are similar to the given image sample
  - Compare the feature vector (signature) extracted from the sample with the feature vectors of images that have already been extracted and indexed in the image database
- Image feature specification queries
  - Specify or sketch image features like color, texture, or shape, which are translated into a feature vector
  - Match the feature vector with the feature vectors of the images in the database

Approaches Based on Image Signature

- **Color histogram-based signature**
  - The signature includes color histograms based on color composition of an image regardless of its scale or orientation
  - No information about shape, location, or texture
  - Two images with similar color composition may contain very different shapes or textures, and thus could be completely unrelated in semantics

- **Multi-feature composed signature**
  - Signature of an image includes composition of multiple features: color, shape, location, and texture.
  - The extracted image features are stored as metadata and images are indexed based on such metadata.
  - Separate distance measures can be defined for each feature and subsequently combined to define overall results.

- **Wavelet-based signature**
  - Use the dominant wavelet coefficients of an image as its signature
  - Wavelets capture shape, texture, and location information in a single unified framework
  - Since this method computes a single signature for an entire image; it may fail to identify images containing similar objects that are in different locations.

- **Wavelet-based signature with region-based granularity**
  - The computation and comparison of signatures are at the granularity of regions ,not the entire image.
  - This is based on the observation that similar images may contain similar regions, but a region in one image could be translation or scaling of a matching region in the other.

Multidimensional Analysis of Multimedia Data

- Multimedia data cube
  - Design and construction similar to that of traditional data cubes from relational data
  - Contain additional dimensions and measures for multimedia information, such as color, texture, and shape
- The database does not store images but their descriptors
  - Feature descriptor: a set of vectors for each visual characteristic
    - Color vector: contains the color histogram
    - MFC (Most Frequent Color) vector: five color centroids
    - MFO (Most Frequent Orientation) vector: five edge orientation centroids
  - Layout descriptor: contains a color layout vector and an edge layout vector

Mining Associations in Multimedia Data

- Association between image and non-image content features
  - "If it atleast 50% of the upper part of the picture is blue, then it is likely to represent sky"
- Association among image content that are not related to spatial relationship
  - If a picture contains two blue squares, then it is likely to contain one red circle as well.
- Associations among image content related to spatial relationship
  - If a red triangle is between two yellow squares, then it is likely a big oval-shaped object is underneath

Mining Text and Web Data

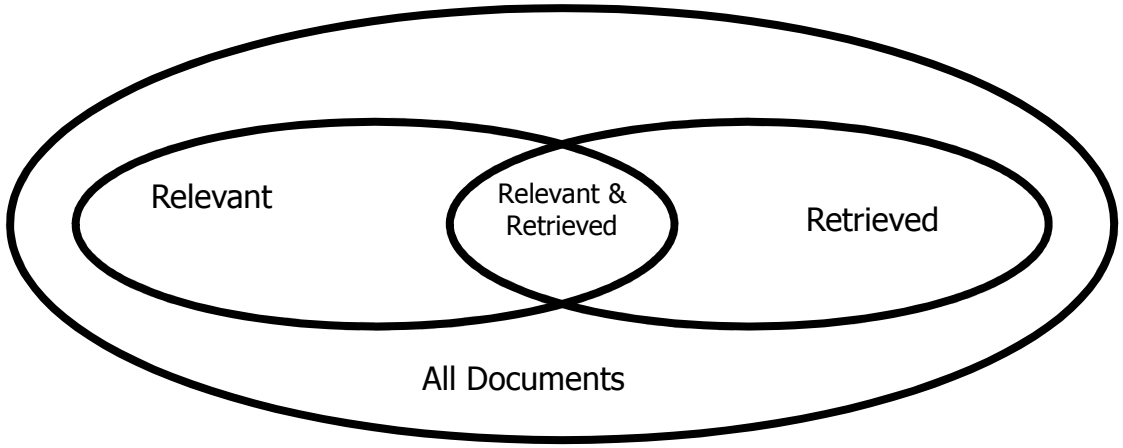
Text Databases and IR

- Text databases (document databases)
  - Large collections of documents from various sources: news articles, research papers, books, digital libraries, e-mail messages, and Web pages, library database, etc.
  - Data stored is usually *semi-structured*
  - Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data
- Information retrieval
  - A field developed in parallel with database systems
  - Information is organized into (a large number of) documents
  - Information retrieval problem: locating relevant documents based on user input, such as keywords or example documents

Information Retrieval

- Typical IR systems
  - Online library catalogs
  - Online document management systems
- Information retrieval vs. database systems
  - Some DB problems are not present in IR, e.g., update, transaction management, complex objects
  - Some IR problems are not addressed well in DBMS, e.g., unstructured documents, approximate search using keywords and relevance

Basic Measures for Text Retrieval



- Precision: the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses)

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- Recall: the percentage of documents that are relevant to the query and were, in fact, retrieved

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Information Retrieval Techniques

- Basic Concepts
  - A document can be described by a set of representative keywords called index terms.
  - Different index terms have varying relevance when used to describe document contents.
  - This effect is captured through the assignment of numerical weights to each index term of a document. (e.g.: frequency, tf-idf)
- DBMS Analogy
  - Index Terms → Attributes
  - Weights → Attribute Values

Information Retrieval Techniques

- Index Terms (Attribute) Selection:
  - Stop list
  - Word stem
  - Index terms weighting methods
- Terms ✕ Documents Frequency Matrices
- Information Retrieval Models:
  - Boolean Model
  - Vector Model
  - Probabilistic Model

Boolean Model

- Consider that index terms are either present or absent in a document
- As a result, the index term weights are assumed to be all binaries
- A query is composed of index terms linked by three connectives: not, and, and or
  - e.g.: car *and* repair, plane *or* airplane
- The Boolean model predicts that each document is either relevant or non-relevant based on the match of a document to the query

Keyword-Based Retrieval

- A document is represented by a string, which can be identified by a set of keywords
- Queries may use expressions of keywords
  - E.g., car *and* repair shop, tea *or* coffee, DBMS *but not* Oracle
  - Queries and retrieval should consider synonyms, e.g., repair and maintenance
- Major difficulties of the model
  - Synonymy: A keyword *T* does not appear anywhere in the document, even though the document is closely related to *T*, e.g., data mining
  - Polysemy: The same keyword may mean different things in different contexts, e.g., mining

Similarity-Based Retrieval in Text Data

- Finds similar documents based on a set of common keywords
- Answer should be based on the degree of relevance based on the nearness of the keywords, relative frequency of the keywords, etc.
- Basic techniques
- Stop list
  - Set of words that are deemed "irrelevant", even though they may appear frequently
  - E.g., *a, the, of, for, to, with*, etc.
  - Stop lists may vary when document set varies

Similarity-Based Retrieval in Text Data

- Word stem
  - Several words are small syntactic variants of each other since they share a common word stem
  - E.g., *drug*, *drugs*, *drugged*
- A term frequency table
  - Each entry *frequent_table(i, j)* = # of occurrences of the word t_i in document d_i
  - Usually, the *ratio* instead of the absolute number of occurrences is used
- Similarity metrics: measure the closeness of a document to a query (a set of keywords)
  - Relative term occurrences
  - Cosine distance:

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|}$$

Types of Text Data Mining

- Keyword-based association analysis
- Automatic document classification
- Similarity detection
  - Cluster documents by a common author
  - Cluster documents containing information from a common source
- Link analysis: unusual correlation between entities
- Sequence analysis: predicting a recurring event
- Anomaly detection: find information that violates usual patterns
- Hypertext analysis
  - Patterns in anchors/links
    - Anchor text correlations with linked objects

Text Classification

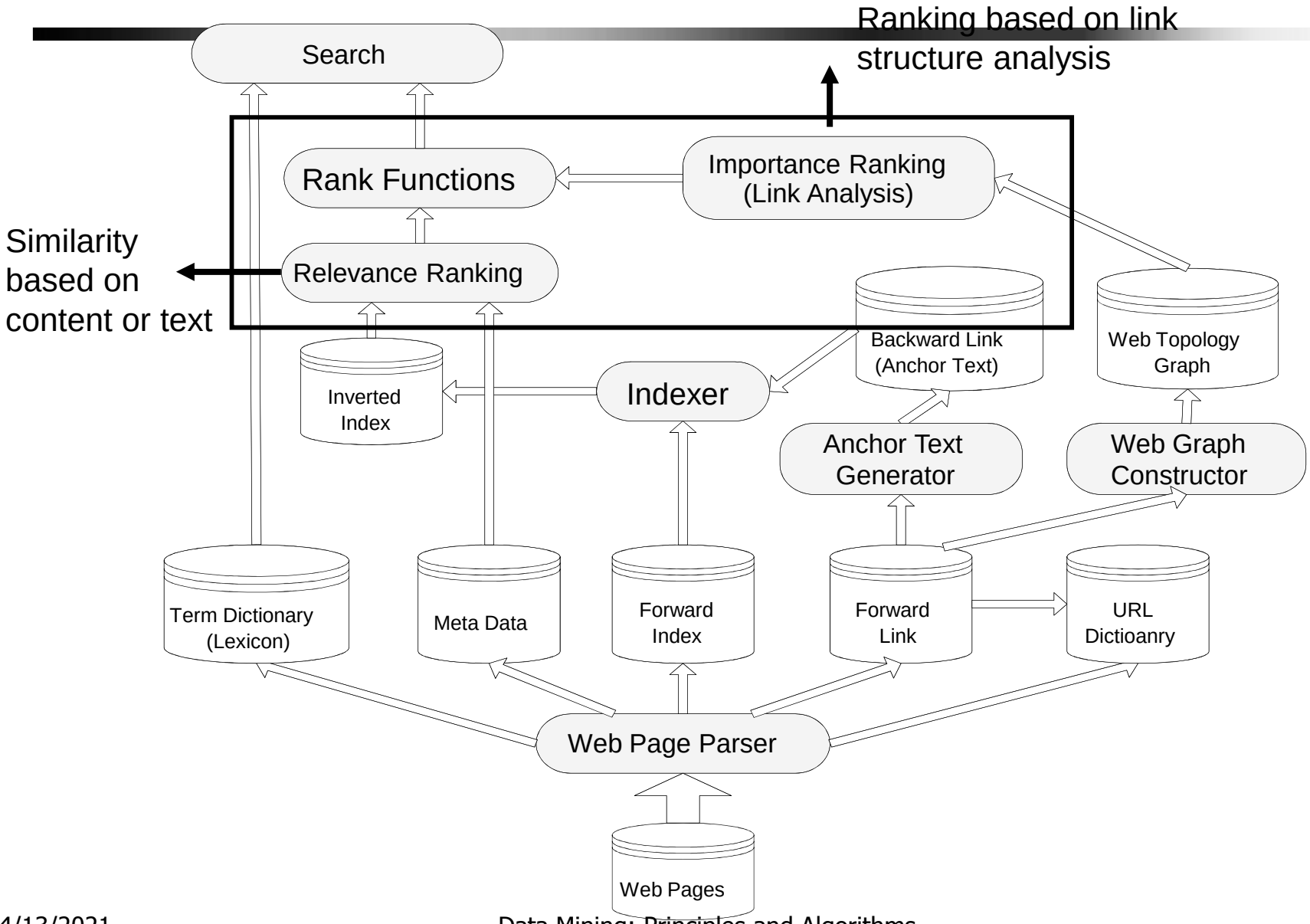
- Motivation
  - Automatic classification for the large number of on-line text documents (Web pages, e-mails, corporate intranets, etc.)
- Classification Process
  - Data preprocessing
  - Definition of training set and test sets
  - Creation of the classification model using the selected classification algorithm
  - Classification model validation
  - Classification of new/unknown text documents
- Text document classification differs from the classification of relational data
  - Document databases are not structured according to attribute-value pairs

Text Classification(2)

- Classification Algorithms:
  - Support Vector Machines
  - K-Nearest Neighbors
  - Naïve Bayes
  - Neural Networks
  - Decision Trees
  - Association rule-based
  - Boosting

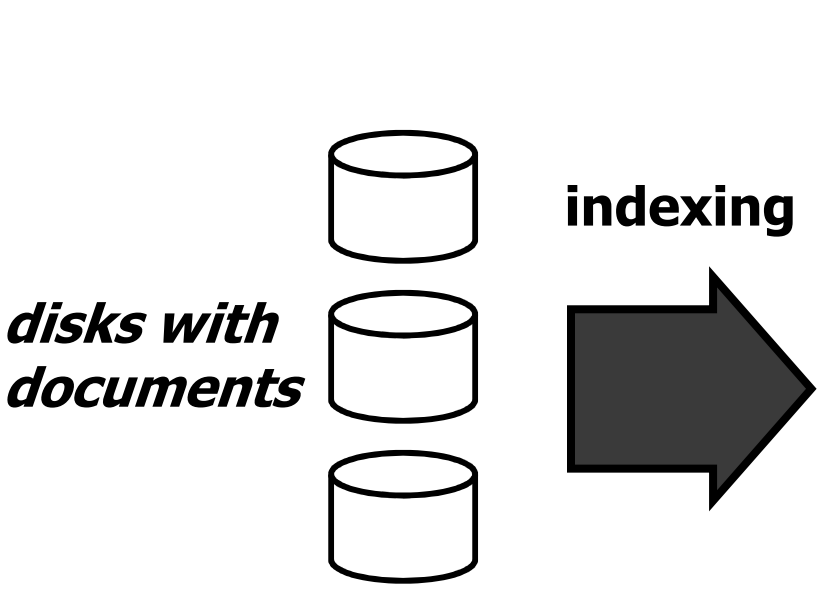
Mining Web Data

Search Engine – Two Rank Functions



Relevance Ranking

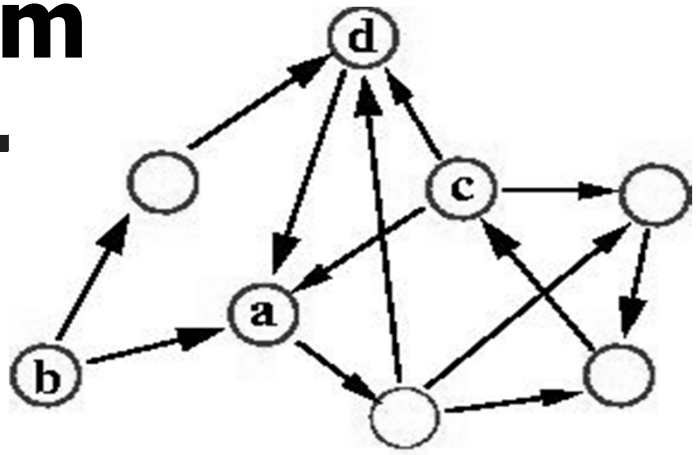
- **Inverted index**
  - **A data structure for supporting text queries**
  - **like index in a book**



aalborg	3452, 11437, …..
.	
.	
.	
.	
.	
arm	4, 19, 29, 98, 143, ...
armada	145, 457, 789, ...
armadillo	678, 2134, 3970, ...
armani	90, 256, 372, 511, ...
.	
.	
.	
.	
.	
zz	602, 1189, 3209, ...

inverted index

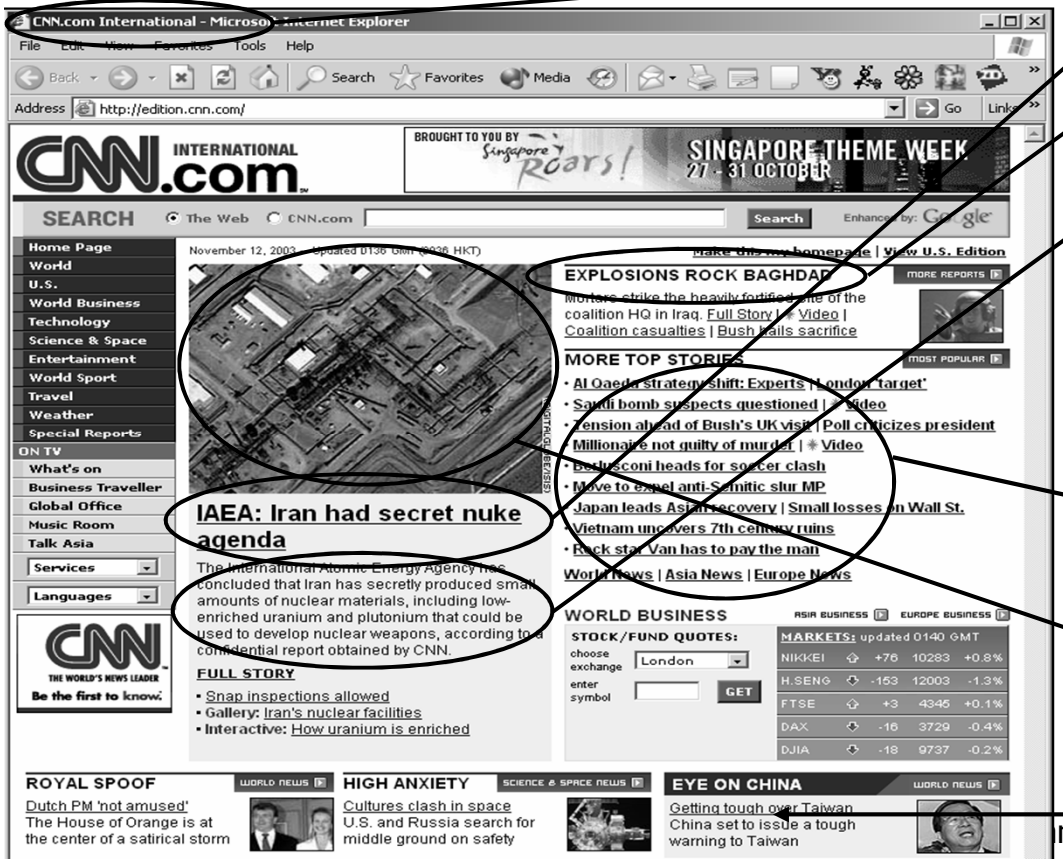
The PageRank Algorithm



- Basic idea
  - ***significance of a page is determined by the significance of the pages linking to it***

Layout Structure

- Compared to plain text, a web page is a 2D presentation
  - Rich visual effects created by different term types, formats, separators, blank areas, colors, pictures, etc
  - Different parts of a page are not equally important



Title: CNN.com International

H1: IAEA: Iran had secret nuke agenda

H3: EXPLOSIONS ROCK BAGHDAD

…

TEXT BODY (with position and font type): The International Atomic Energy Agency has concluded that Iran has secretly produced small amounts of nuclear materials including low enriched uranium and plutonium that could be used to develop nuclear weapons according to a confidential report obtained by CNN…

Hyperlink:

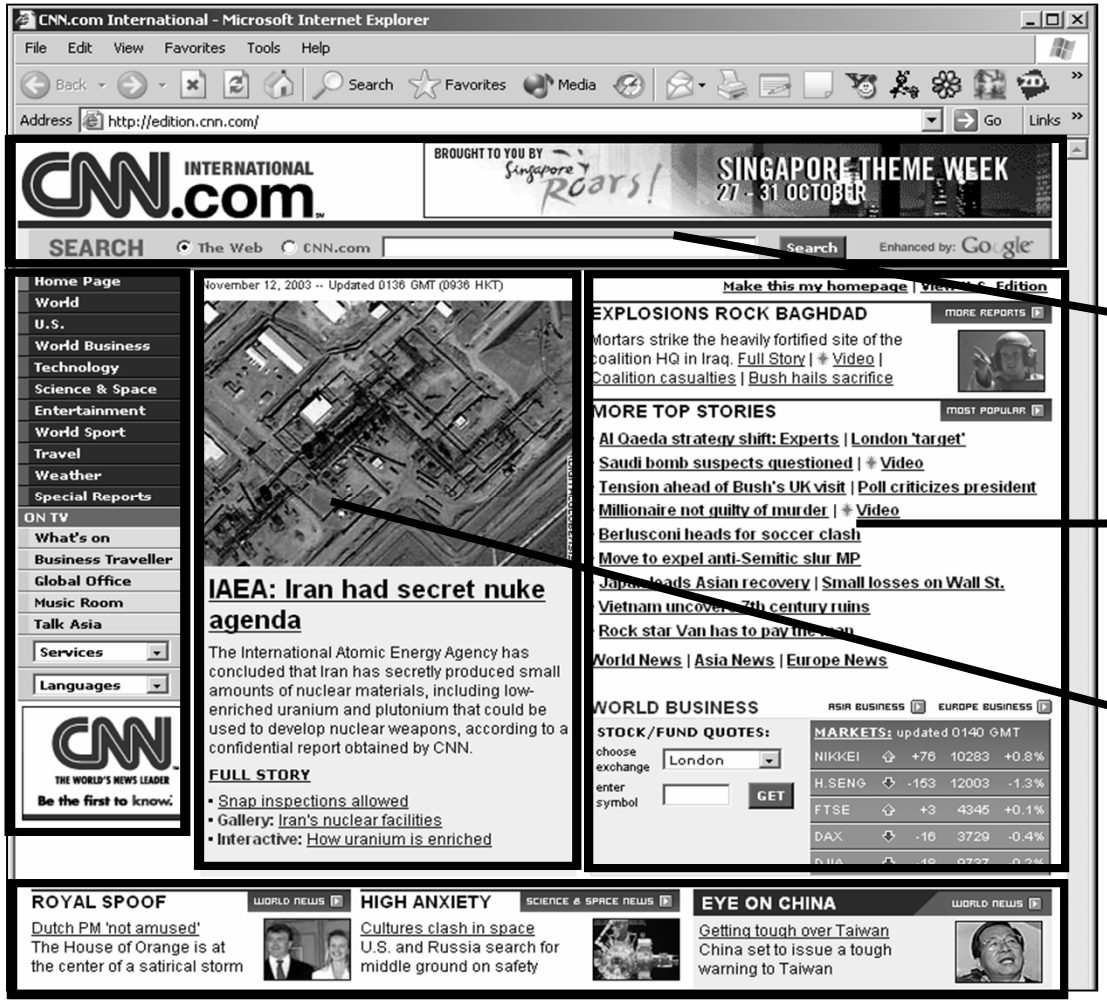
- URL: http://www.cnn.com/...
- Anchor Text: Al oaeda…

Image:

- URL: http://www.cnn.com/image/...
- Alt & Caption: Iran nuclear …

Anchor Text: CNN Homepage News …

Web Page Block—Better Information Unit



Web Page Blocks

Importance = Low

Importance = Med

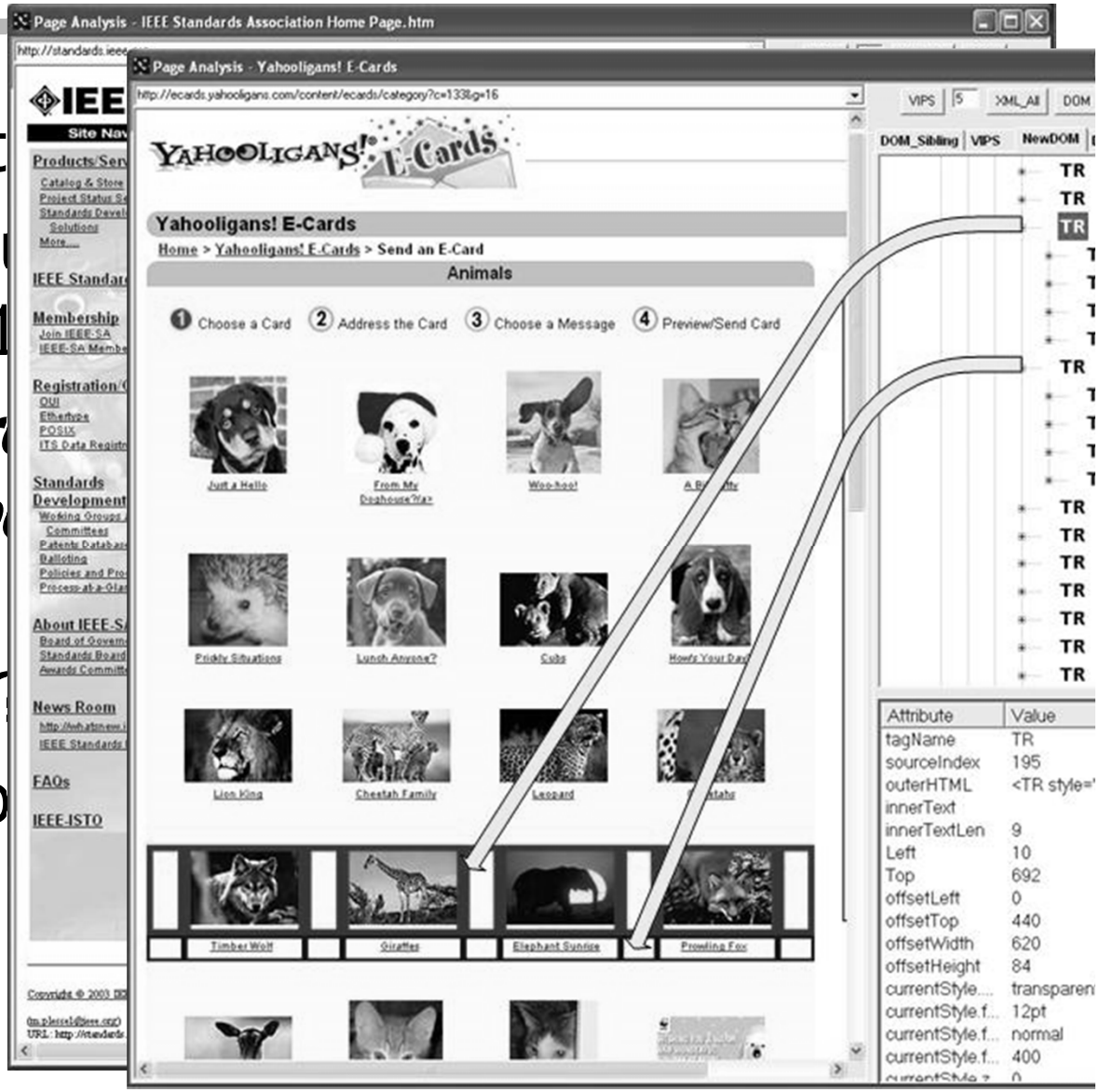
Importance = High

Motivation for VIPS (VIsion-based Page Segmentation)

- Problems of treating a web page as an atomic unit
  - Web page usually contains not only pure content
    - Noise: navigation, decoration, interaction, …
  - Multiple topics
  - Different parts of a page are not equally important
- Web page has internal structure
  - Two-dimension logical structure & Visual layout presentation
  - **>** Free text document
  - **<** Structured document
- Layout – the 3rd dimension of Web page
  - 1st dimension: content
  - 2nd dimension: hyperlink

Is DOM a Good Representation of Page Structure?

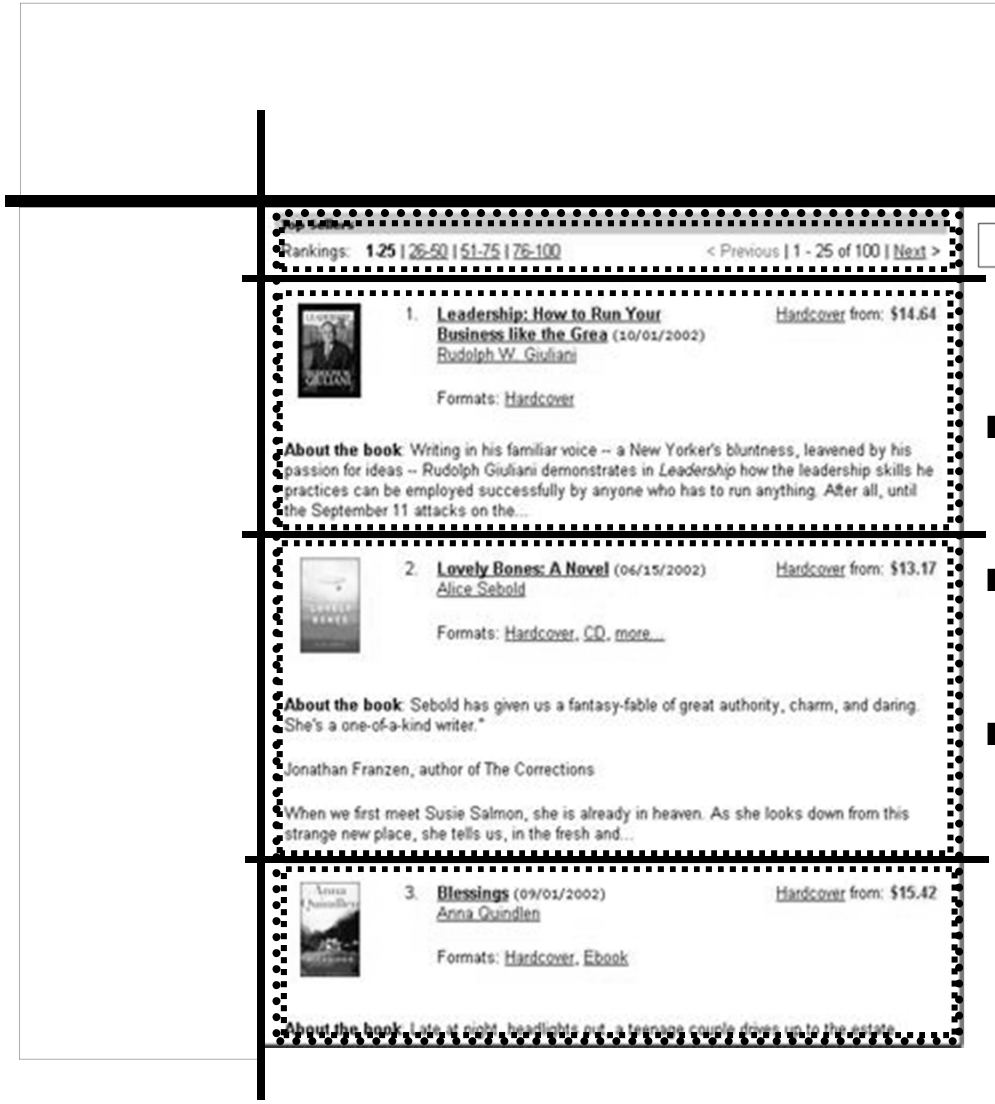
- Page segmentat
  - Extract structu
    UL, TITLE, H1
  - ***DOM is mor***
    does not ne
    structure
- How about XML
  - A long way to

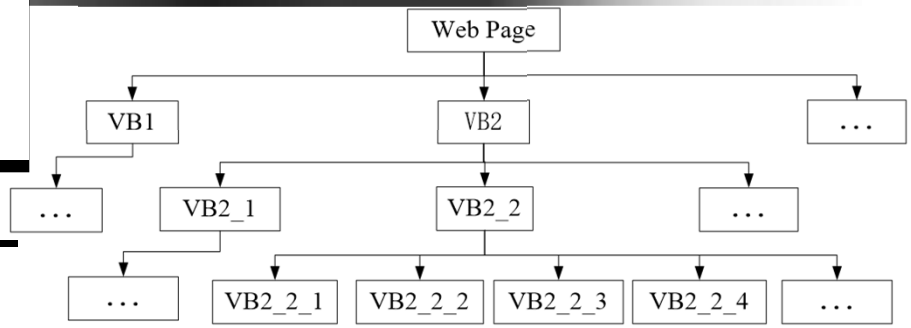


VIPS Algorithm

- Motivation:
  - In many cases, topics can be distinguished with visual clues. Such as position, distance, font, color, etc.
- Goal:
  - Extract the semantic structure of a web page based on its visual presentation.
- Procedure:
  - Top-down partition the web page based on the separators
- Result
  - A tree structure, each node in the tree corresponds to a block in the page.
  - Each node will be assigned a value (Degree of Coherence) to indicate how coherent of the content in the block based on visual perception.
  - Each block will be assigned an importance value
  - Hierarchy or flat

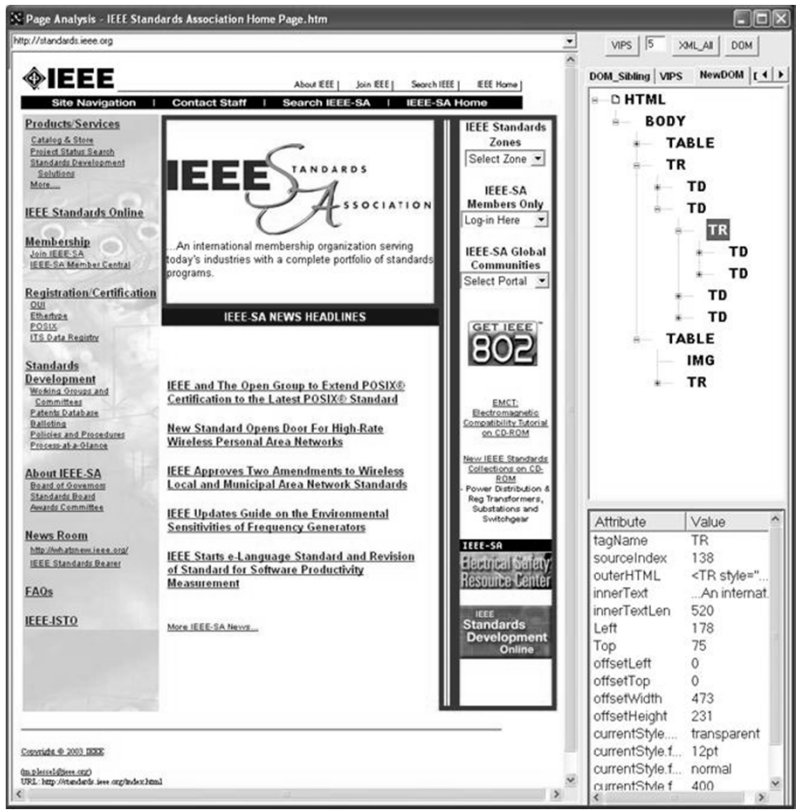
VIPS: An Example



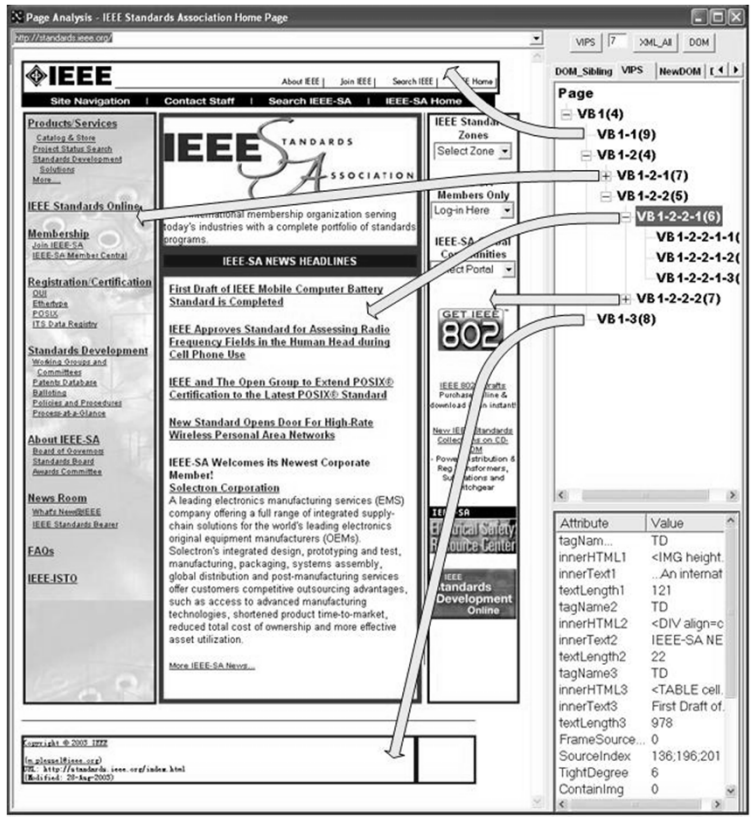


- A hierarchical structure of layout block
- A Degree of Coherence (DOC) is defined for each block
- The Permitted Degree of Coherence (PDOC) can be pre-defined to achieve different granularities for the content structure

Example of Web Page Segmentation (1)

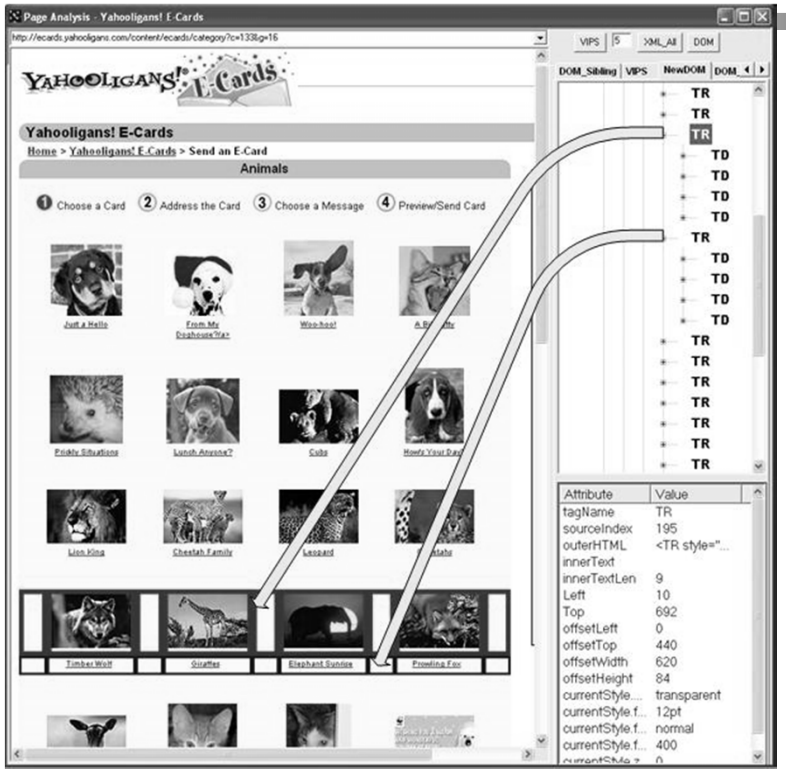


(DOM Structure)



(VIPS Structure)

Example of Web Page Segmentation (2)



(DOM Structure)

(VIPS Structure)

- Can be applied on web image retrieval

Hierarchical Clustering

- Clustering based on three representations
  - Visual feature
    - Hard to reflect the semantic meaning
  - Textual feature
    - Semantic
    - Sometimes the surrounding text is too little
  - Link graph:
    - Semantic
    - Many disconnected sub-graph (too many clusters)
- Two Steps:
  - Using texts and link information to get semantic clusters
  - For each cluster, using visual feature to re-organize the images to facilitate user's browsing

Clustering Using Visual Feature

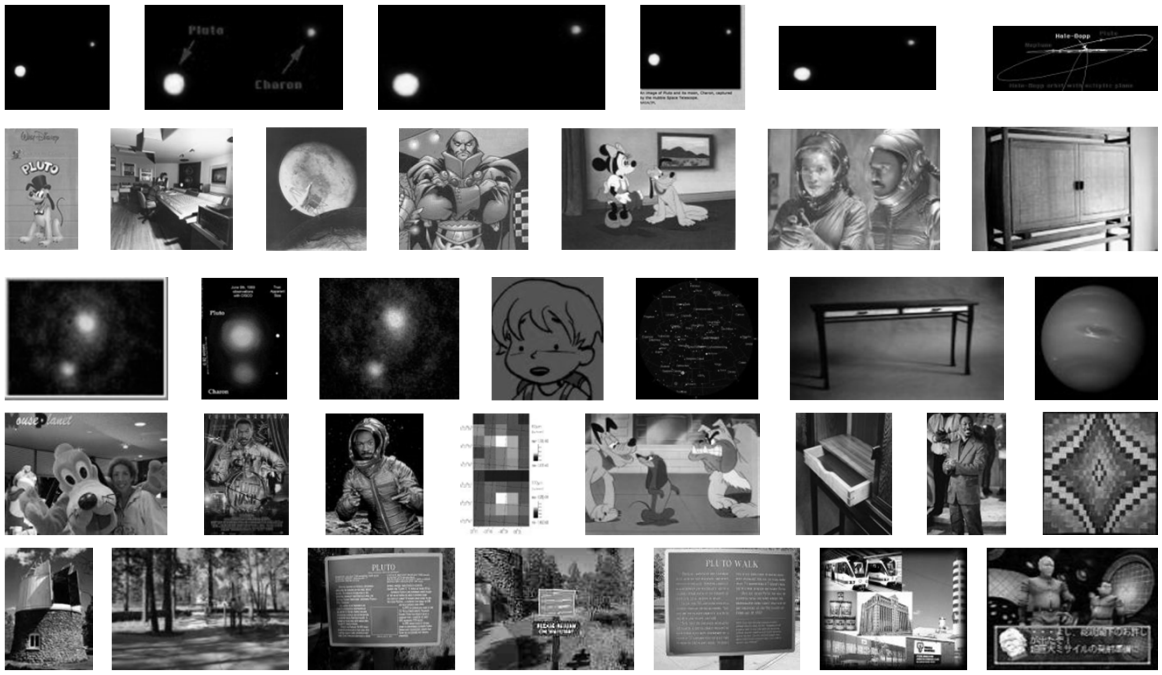


Figure 5. Five clusters of search results of query "pluto" using low level visual feature. Each row is a cluster.

- From the perspectives of color and texture, the clustering results are quite good. Different clusters have different colors and textures. However, from semantic perspective, these clusters make little sense.

Clustering Using Textual Feature

Figure 7. Six clusters of search results of query "pluto" using textual feature. Each row is a cluster

- Six semantic categories are correctly identified if we choose $k = 6$.

Clustering Using Graph Based Representation

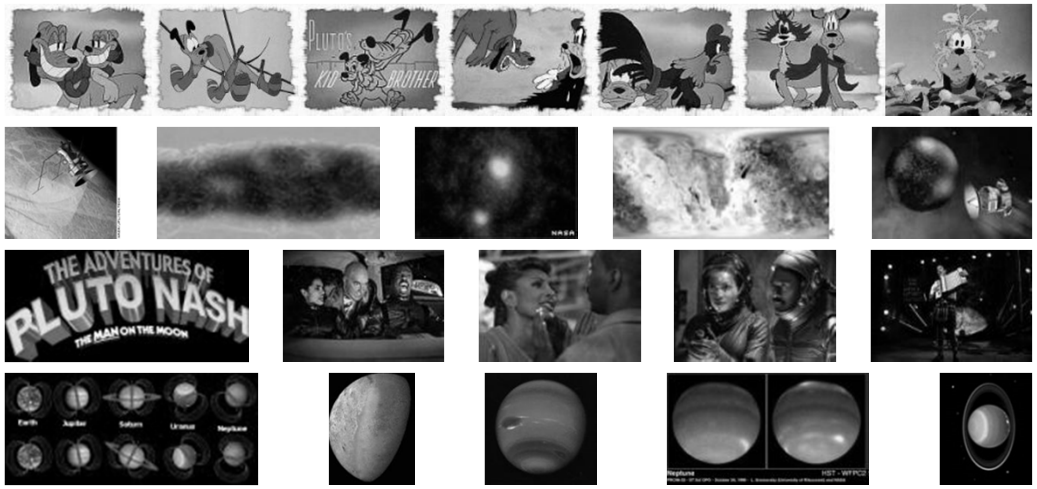


Figure 8. Five clusters of search results of query "pluto" using image link graph. Each row is a cluster

- Each cluster is semantically aggregated.
- Too many clusters.

All The Best