Multimedia Mining

— edited by Manjunath —

©Jiawei Han and Micheline Kamber
Intelligent Database Systems Research Lab
School of Computing Science
Simon Fraser University, Canada
http://www.cs.sfu.ca

Generalizing Spatial and Multimedia Data

- Spatial data:
  - Generalize detailed geographic points into clustered regions, such as business, residential, industrial, or agricultural areas, according to land usage
  - Require the merge of a set of geographic areas by spatial operations
- Image data:
  - Extracted by aggregation and/or approximation
  - Size, color, shape, texture, orientation, and relative positions and structures of the contained objects or regions in the image
- Music data:
  - Summarize its melody: based on the approximate patterns that repeatedly occur in the segment
  - Summarized its style: based on its tone, tempo, or the major musical instruments played
- Text Data:
  - Text document retrieval, key-word search and indexing
  - Cluster doucuments

Mining Complex Types of Data

- Mining text databases
- Content based access to image/video databases
- Relevance Feedback
- Summary

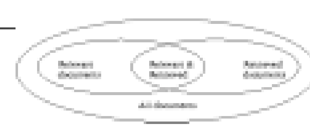
Text Databases and IR

- 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

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

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 Databases

- 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, with*, etc.
  - Stop lists may vary when document set varies

Similarity-Based Retrieval in Text Databases (2)

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

Document term matrix:example

	T1	T2	T3	T4	T5	T6
D1	24	21	9	0	0	3
D2	32	10	5	0	3	0
D3	12	16	5	0	0	0
D4	6	7	2	0	0	0
D5	43	31	20	0	3	0
D6	2	0	0	18	7	16
D7	0	0	1	32	12	0
D8	3	0	0	22	4	2
D9	1	0	0	34	27	25
D10	6	0	0	17	4	23

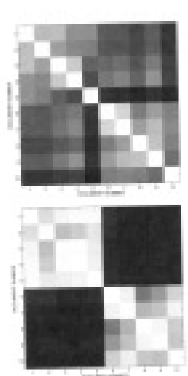
T1=database
T2=sql
T3=index
T4=regression
T5=likelihood
T6=linear

M=10x6 matrix

Figure 14.2 from Hand's book

Euclidean distance

Cosine distance



Latent Semantic Indexing

- Basic idea
  - Similar documents have similar word frequencies
  - Difficulty: the size of the term frequency matrix is very large
  - Use a singular value decomposition (SVD) techniques to reduce the size of frequency table
  - Retain the *K* most significant rows of the frequency table
- Method
  - Create a term frequency matrix, *freq_matrix*
  - SVD construction: Compute the singular valued decomposition of *freq_matrix* by splitting it into 3 matrices, *U*, *S*, *V*
  - Vector identification: For each document *d*, replace its original document vector by a new excluding the eliminated terms
  - Index creation: Store the set of all vectors, indexed by one of a number of techniques (such as TV-tree)

LSI

- M= U S V^T
- U = 10 x 6 matrix where each row is a vector of weights for a particular document
- S = 6 x 6 diagonal matrix of eigenvalues for each principal component direction
- V^T = 6 x 6 matrix, columns of which are the new basis
- For the previous example, the diagonal elements of S = { 77 69 23 14 12 5}
  - The first two principal component directions contain 92.5% of the "energy".

Reduced components

D1	30.9	-11.5
D2	30.3	-10.8
D3	18	-7.7
D4	8.4	-3.6
D5	52.7	-20.7
D6	14.2	21.8
D7	10.8	21.9
D8	11.5	28
D9	9.5	17.8
D10	19.9	45.1

V1= [0.74 0.49 0.27 0.28 0.18 0.19]

V2=[-0.3 -0.24 -0.12 0.74 0.37 0.31]

These are the two directions having the maximum variance of the data

Note that d1 (database) and d2 (SQL) are very similar in the SVD space.

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

Keyword-based association analysis

- Collect sets of keywords or terms that occur frequently together and then find the association or correlation relationships among them
- First preprocess the text data by parsing, stemming, removing stop words, etc.
- Then evoke association mining algorithms
  - Consider each document as a transaction
  - View a set of keywords in the document as a set of items in the transaction
- Term level association mining
  - No need for human effort in tagging documents
  - The number of meaningless results and the execution time is greatly reduced

Automatic document classification

- Motivation
  - Automatic classification for the tremendous number of on-line text documents (Web pages, e-mails, etc.)
- A classification problem
  - Training set: Human experts generate a training data set
  - Classification: The computer system discovers the classification rules
  - Application: The discovered rules can be applied to classify new/unknown documents
- Text document classification differs from the classification of relational data
  - Document databases are not structured according to attribute-value pairs

Association-Based Document Classification

- Extract keywords and terms by information retrieval and simple association analysis techniques
- Obtain concept hierarchies of keywords and terms using
  - Available term classes, such as WordNet
  - Expert knowledge
  - Some keyword classification systems
- Classify documents in the training set into class hierarchies
- Apply term association mining method to discover sets of associated terms
- Use the terms to maximally distinguish one class of documents from others
- Derive a set of association rules associated with each document class
- Order the classification rules based on their occurrence frequency and discriminative power
- Used the rules to classify new documents

Document Clustering

- Automatically group related documents based on their contents
- Require no training sets or predetermined taxonomies, generate a taxonomy at runtime
- Major steps
  - Preprocessing
    - Remove stop words, stem, feature extraction, lexical analysis, …
  - Hierarchical clustering
    - Compute similarities applying clustering algorithms, …
  - Slicing
    - Fan out controls, flatten the tree to configurable number of levels, …

Mining Complex Types of Data

- Mining text databases
- Content Based Image/Video Retrieval
- Relevance Feedback
- Summary

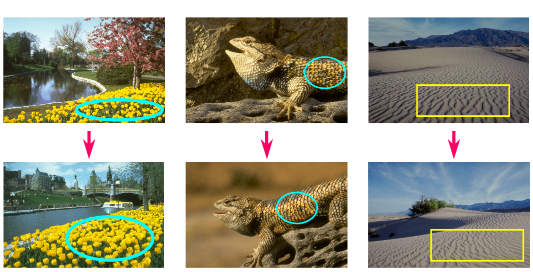
Similarity Search in Multimedia Data

- Description-based retrieval systems
  - Build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation
  - Labor-intensive if performed manually
  - Results are typically of poor quality if automated
- 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:
  - Query by example: 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

Query by Example



©Salembier

MPEG-7 standard for Image/Video Representation

- **MPEG-1**: Storage of moving picture and audio on storage media (CD-ROM) *11 / 1992*
- **MPEG-2**: Digital television *11 / 1994*
- **MPEG-4**: Coding of natural and synthetic media objects for multimedia applications *v1: 09 / 1998* *v2: 11 / 1999*
- **MPEG-7**: Multimedia content description for AV material *08 / 2001*
- **MPEG-21**: Digital audiovisual framework: Integration of multimedia technologies (identification, copyright, protection, etc.) *11 / 2001*

©Salembier

Objective of MPEG-7

- Standardize content-based description for various types of audiovisual information
  - Enable fast and efficient content searching, filtering and identification
  - Describe several aspects of the content (low-level features, structure, semantic, models, collections, creation, etc.)
  - Address a large range of applications (⇒ user preferences)
- Types of audiovisual information:
  - Audio, speech
  - Moving video, still pictures, graphics, 3D models
  - Information on how objects are combined in scenes
- Descriptions independent of the data support
- Existing solutions for textual content or description

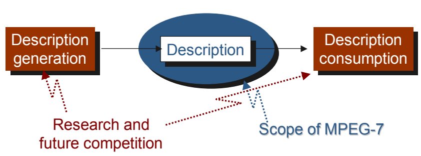
©Salembier

Example of application areas

- Storage and retrieval of audiovisual databases (image, film, radio archives)
- Broadcast media selection (radio, TV programs)
- Surveillance(traffic control, surface transportation, production chains)
- E-commerce and Tele-shopping (searching for clothes / patterns)
- Remote sensing(cartography, ecology, natural resources management)
- Entertainment (searching for a game, for a karaoke)
- Cultural services (museums, art galleries)
- Journalism (searching for events, persons)
- Personalized news service on Internet (push media filtering)
- Intelligent multimedia presentations
- Educational applications
- Bio-medical applications

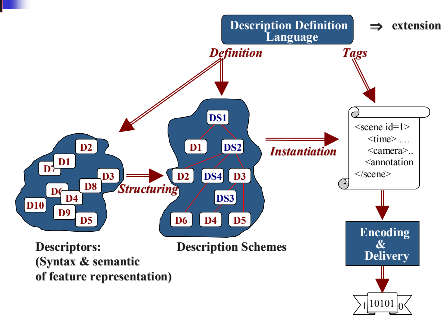
©Salembier

Scope of MPEG-7



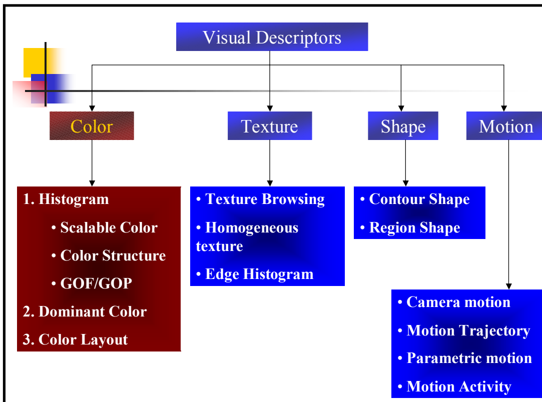
- **The description generation**
  - Feature extraction, Indexing process, Annotation &Authoring tools,...)
- **consumption**
  - Search engine, Filtering tool, Retrieval process, Browsing device, ...)
- **are non normative parts of MPEG-7**
- The goal is to define the minimum that enables interoperability

MPEG-7 working areas



Parts of the MPEG-7 Standard

- ISO / IEC 15938 - 1: Systems
- ISO / IEC 15938 - 2: Description Definition Language
- ISO / IEC 15938 - 3: Visual
- ISO / IEC 15938 - 4: Audio
- ISO / IEC 15938 - 5: Multimedia Description Schemes
- ISO / IEC 15938 - 6: Reference Software



Performance evaluation

- Let the number of ground truth images for a query q be NG(q)
- Compute NR(q), number of found items in first K retrievals, where
- - K=min(4*NG(q), 2*GTM)
  Where GTM is max{NG(q)} for all q's of a data set.
- - Compute MR(q)=NG(q)-NR(q), number of missed items
- - Compute from the ranks *Rank*(*k*) of the found items counting the rank of the first retrieved item as one.
- - A Rank of (**K+1**) is assigned to each of the ground truth images which are not in the first K retrievals.
- - Compute the normalized modified retrieval rank as follows (next slide). Note that *the NMRR*(*q*) will always be in the range of [0.0,1.0].

Average Retrieval Rate (AVR) and ANMRR

Compute AVR(q) for query q as follows:

$$AVR(q) = \sum_{k=1}^{NG(q)} \frac{Rank(k)}{NG(q)}$$

Compute the modified retrieval rank as follows:

$$MRR(q) = AVR(q) - 0.5 - \frac{NG(q)}{2}$$

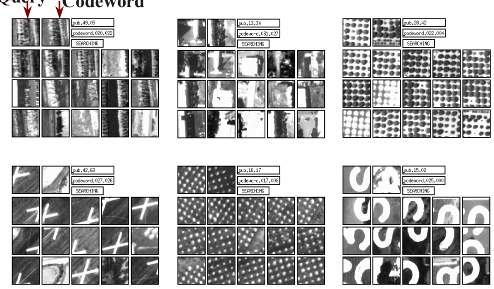
Normalized MRR, NMRR = MRR(q)/Norm(q)

Where Norm(q)=1.25*K –0.5 – 0.5*NG(q)

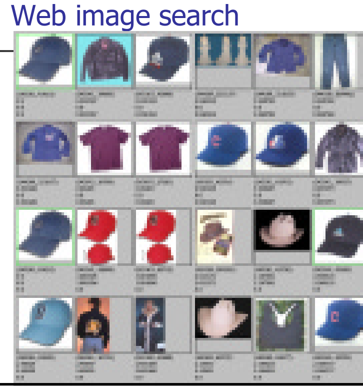
$$ANMRR = \frac{1}{Q}\sum_{q=1}^{Q} NMRR(q)$$

Texture based similarity search

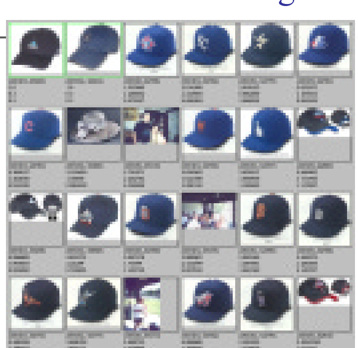
Query **Codeword**



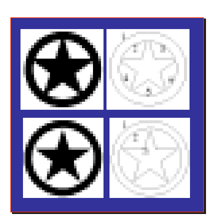
Web image search



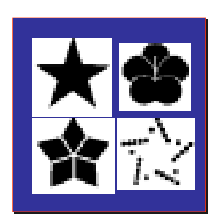
Applications - Web Image Search



Shape Descriptors



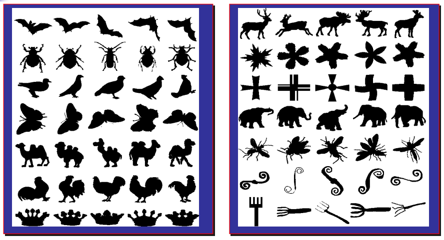
Contour-based shape descriptor



Region-based shape descriptor

Experimental Dataset & Procedure

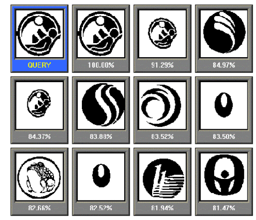
1/3



- *Dataset 1*: 70 classes × 20 variations = 1400 images
- CE1-A-1: Scale, CE1-A-2: Rotation, CE1-B: Similarity

Retrieval Example

1/2



Query results without respect to perspective normalization

Query results with respect to perspective normalization

Mining Associations in Multimedia Data

- Special features:
  - Need # of occurrences besides Boolean existence, e.g.,
    - "Two red square and one blue circle" implies theme "air-show"
  - Need spatial relationships
    - Blue on top of white squared object is associated with brown bottom
  - Need multi-resolution and progressive refinement mining
    - It is expensive to explore detailed associations among objects at high resolution
    - It is crucial to ensure the completeness of search at multi-resolution space

Challenge: Curse of Dimensionality

- Difficult to implement a data cube efficiently given a large number of dimensions, especially serious in the case of multimedia data cubes
- Many of these attributes are set-oriented instead of single-valued
- Restricting number of dimensions may lead to the modeling of an image at a rather rough, limited, and imprecise scale
- More research is needed to strike a balance between efficiency and power of representation

Mining Complex Types of Data

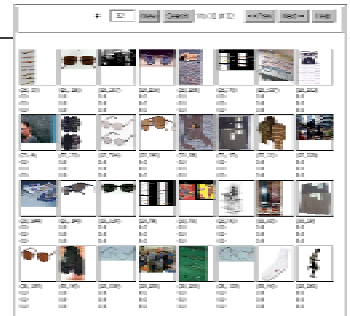
- Mining text databases
- Content based Image/Video Retrieval
- Relevance Feedback
- Summary

Relevance Feedback (RF)

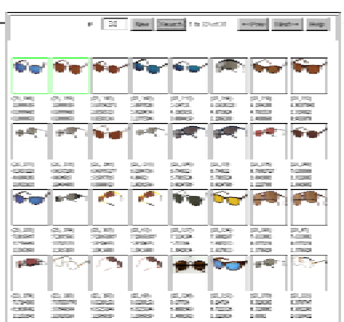
- Low level features do not capture well the semantics
- Relevance feedback:
  - Learning mechanism
  - Learns user's subjective similarity measures
  - Aid to effective high-level concept query
- Relevance Feedback application for large datasets
  - Web
    - Online image search
  - Scientific image repositories
    - Geography - Aerial
    - Medical - Neuroscience

Large Image Dataset Search

- Search semantically diverse image datasets
- Approximate search process:
  - Keyword search to identify likely categories
  - NN search within these categories using color and texture
- User selects query images of interest
- Either the query or the similarity measure or both are updated
- EXAMPLE:
  - The keyword “sunglasses” identifies ‘Eyewear’ category
  - 32 random images from the 458 images in these category



After two RF iterations



Relevance Feedback Implementation

- Weight Matrix update
  - User’s feedback: $W_{t-1} \rightarrow W_t$
  - NN search
    - Similarity measure: $d(Q,F_i,W_t) = (Q-F_i)^T W_t (Q-F_i)$
  - Efficient, linear mapping
- Kernel-based learning
  - non-linear mapping $\phi(F_i)$
  - NN search in the mapped space: $F_i \rightarrow \phi(F_i)$
    - Similarity measure:

$$d(\phi(Q),\phi(F_i)) = (\phi(Q)-\phi(F_i))^T(\phi(Q)-\phi(F_i))$$

  - Computationally expensive, more effective

Weight Matrix Update
Weighted Euclidean Distance

- K' - number of identified relevant objects
- $X_k = [x_{k1}, x_{k2}, \ldots x_{kM}]$ - relevant vectors
- σ_m - standard deviation of the sequence of
- Diagonal weight matrix update (MARS):

$$(W_t)_m = \frac{(\prod_{i=1}^{M} \sigma_i^2)^{\frac{1}{M}}}{\sigma_m^2}$$

- Distance measure: $d(Q,F_i,W_t) = (Q-F_i)^T W_t (Q-F_i)$

Weight Matrix Update
Quadratic Distance

- Weight matrix: $W_t = P_t^T \lambda_t P_t$
  - real, symmetric, positive definite
- Mahalanobis distance update (MindReader):
  - Sample covariance matrix: $C = P_C^T \lambda_C P_C$

$$(C)_{ij} = \frac{\sum_{k=1}^{K'} \pi_k (x_{ki} - q_i)(x_{kj} - q_j)}{\sum_{k=1}^{K'} \pi_k}$$

  - Weight matrix update:

$$W_t = (\det(C))^{\frac{1}{M}} C^{-1} \rightarrow P_t = P_C, (\lambda_t)_m = \frac{\prod_{i=1}^{M} (\lambda_C)_i}{(\lambda_C)_m}$$

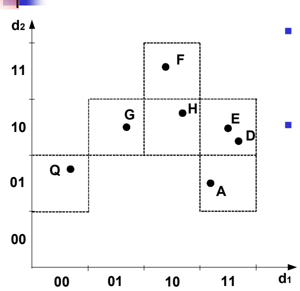
Nearest Neighbor (NN) Computation

- Set of nearest neighbors is computed for each iteration.
- Index should support:
  - Efficient search of large high-dimensional feature set
  - NN queries for all RF scenarios
- Compression based techniques are suitable:
  - Scalar quantization approach - VA-file
  - Vector quantization approach – VQ-file

Paper: Nearest Neighbor Search for Relevance feedback

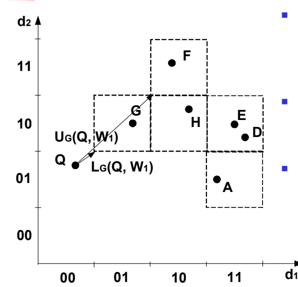
- By Jelena and Manjunath, CVPR 2003.
- Will make the paper available on line

Construction of VA-File



- Feature vector dimension is partitioned into uniform non-overlapping segment
- The approximation for feature vector A is "1101"
  - 11 - index of dimension 1
  - 01 - index of dimension 2

K-NN search VA-File

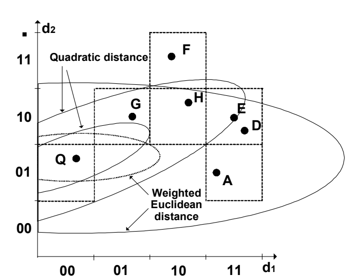


- Two phase search:
  - Phase I - approximation level filtering
  - Phase 2- data level filtering
- K=2
  - Phase 1: N1={A, D, E, F, G, H}
  - Phase 2: R1={G, H}
- Objective:
  minimize set of candidates that contains all the K nearest neighbors

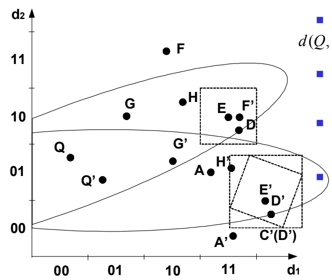
K-NN Search in VA-File

- Bounds: $L_i(Q,W_t) \le d(Q,F_i,W_t) \le U_i(Q,W_t)$
- Two phases filtering
  - Phase I filtering (approximation level): if the lower bound is larger than the K-th largest upper bound encountered so far, skip the approximation (N1 candidates)
  - Phase II filtering (data level): visit the N1 feature vectors in the increasing order of their lower bounds. If a lower bound is larger than the K-th largest actual distance encountered so far, skip the rest of candidates (N2 feature vectors)

Bound Computation in Relevance Feedback



Bound Computation Quadratic Distance



- Distance:
  $d(Q,F_i,W_t)=(Q-F_i)^T W_t (Q-F_i)$
- Weight Matrix:
  $W_t = P_t^T \lambda_t P_t$
- Mapping:
  $D' = P_t D$
- Quadratic distance is weighted Euclidean distance in the mapped space.

Adaptive NN Search in the Presence of Relevance Feedback

- Approximate level filtering
  - Subset of approximations that contains K nearest neighbors
  - Filtering based on VA lower bounds
  - Introduces false candidates
- Spatial mapping
  - Enables us to use VA-file index
  - Approximates VA lower bound for ellipsoid queries
  - More false candidates due to approximation
- Our approach – exploit the correlation between two consecutive NN sets
  - Firmer Phase I filtering bounds
  - Avoid computing and buffering upper bounds
  - Speed up nearest neighbor search algorithm
  - Supports a changing distance metric

Upper Bound on NN set

- Set of nearest neighbors of query Q at iteration t is defined as: $R_t=\{F_k^t, k\in[1,K]\}$
- Upper bound of that set:
  $$r_t(Q)=\max\{d(Q,\ F_k^t,\ W_t)\}$$
- If t>1, define: $r_t^u(Q)=\max\{d(Q,\ F_k^{t-1},\ W_t)\}$
- Then: $\boxed{r_t(Q)\le r_t^u(Q)}$
- Maximum of K distances between the query Q and objects in R_t computed using W_t can not be larger than $r_t^u(Q)$
  $L_i(Q,W_t)\le d(Q,F_i,W_t)\to$ $\boxed{\max\{L_k^{(t-1)}(Q,W_t)\}\le r_t^u(Q)}$

A tighter bound on NN set

- In similar manner, for t>1, define
  $$l_t^u(Q)=\max\{L_k^{(t)}(Q,\ W_t)\}$$
- Then: $\boxed{\max(L_k^{(t-1)}(Q,W_t))\le l_t^u}$
- Maximum of the lower bounds between the query Q and objects in R_t computed using W_t can not be larger than $l_t^u(Q)$
- Therefore: $L_i(Q,W_t)\le d(Q,F_i,W_t)\to l_t^u(Q)\le r_t^u(Q)$
- Filtering bounds: $\boxed{L_i(Q,W_t)\le l_t^u(Q)\le r_t^u(Q)}$

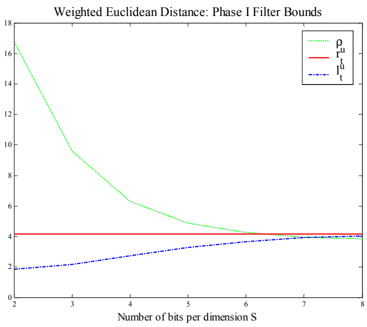
Adaptive NN Search Phase I filtering

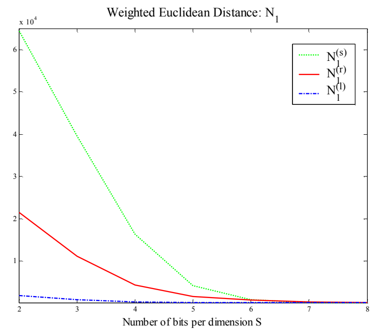
1) if t=1, use the traditional NN search; BREAK;
2) Given R_{t-1} and W_t compute $r_t^u(Q)$ and $l_t^u(Q)$
3) for i=1 to N
   a) Compute $L_i(Q,W_t)$ for weighted distance
   b) Compute $L_i(Q,W_t)\approx L'_i(Q',\lambda_t)$ for quadratic dist.
      If $L_i(Q,W_t)\le r_t^u(Q)$ insert $C(F_i)$ into $N_1^{(r)}(Q,W_t)$
      If $L_i(Q,W_t)\le l_t^u(Q)$ insert $C(F_i)$ into $N_1^{(l)}(Q,W_t)$

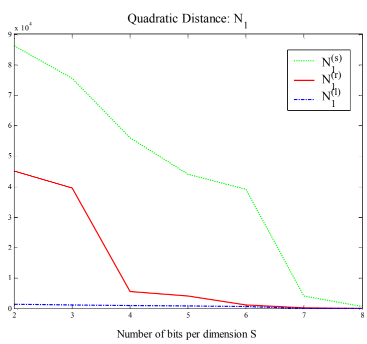
Experiments

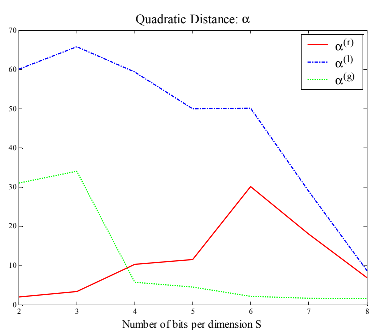
- Dataset on N=90774 feature vectors
- Approximations: constructed at resolution S.
  $$S\in[2,3,4,5,6,7,8]$$
- Queries: $Q_i, i\in[1,\ldots,I]$, I =20, M=60, K'=70.
- Average number of Phase I candidates:
  $$N_1(W_t)=\frac{1}{I}\sum_{i=1}^{I}N_1(Q_i,W_t)$$
- Define effectiveness measures as:

$$\alpha^{(r)}=\frac{1}{I}\sum_{i=1}^{I}\frac{N_1^{(s)}(Q_i,W_t)}{N_1^{(r)}(Q_i,W_t)}\quad \alpha^{(l)}=\frac{1}{I}\sum_{i=1}^{I}\frac{N_1^{(s)}(Q_i,W_t)}{N_1^{(l)}(Q_i,W_t)}\quad \alpha=\frac{1}{I}\sum_{i=1}^{I}\frac{N_1^{(r)}(Q_i,W_t)}{N_1^{(l)}(Q_i,W_t)}$$









Contributions

- An adaptive NN search scheme for relevance feedback:
  - Utilizing the correlation to confine the search space
  - The constraints can be computed efficiently
  - Good bound prediction based on the previous iteration
  - Significant savings on disk accesses
- Work in progress:
  - Approximate search for kernel-based methods

Mining Complex Types of Data

- Mining text databases
- Content based Image/Video Retrieval
- Relevance Feedback
- Summary

Summary (1)

- Mining complex types of data include spatial, multimedia, time-series, text, and Web data
- Object data can be mined by multi-dimensional generalization of complex structured data, such as plan mining for flight sequences
- Spatial data warehousing, OLAP and mining facilitates multidimensional spatial analysis and finding spatial associations, classifications and trends
- Multimedia data mining needs content-based retrieval, similarity search and relevance feedback integrated with mining methods

Summary (2)

- Time-series/sequential data mining includes trend analysis, similarity search in time series, mining sequential patterns and periodicity in time sequence
- Text mining goes beyond keyword-based and similarity-based information retrieval and discovers knowledge from semi-structured data using methods like keyword-based association and document classification
- Web mining includes mining Web link structures to identify authoritative Web pages, the automatic classification of Web documents, building a multilayered Web information base, and Weblog mining

References (1)

- R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. In Proc. 4th Int. Conf. Foundations of Data Organization and Algorithms, Chicago, Oct. 1993.
- R. Agrawal, K.-I. Lin, H.S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. VLDB'95, Zurich, Switzerland, Sept. 1995.
- G. Arocena and A. O. Mendelzon. WebOQL : Restructuring documents, databases, and webs. ICDE'98, Orlando, FL, Feb. 1998.
- R. Agrawal, G. Psaila, E. L. Wimmers, and M. Zait. Querying shapes of histories. VLDB'95, Zurich, Switzerland, Sept. 1995.
- R. Agrawal and R. Srikant. Mining sequential patterns. ICDE'95, Taipei, Taiwan, Mar. 1995.
- S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. WWW'98, Brisbane, Australia, 1998.
- C. Bettini, X. Sean Wang, and S. Jajodia. Mining temporal relationships with multiple granularities in time sequences. Data Engineering Bulletin, 21:32-38, 1998.
- R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley, 1999.
- S. Chakrabarti, B. E. Dom, and P. Indyk. Enhanced hypertext classification using hyper-links. SIGMOD'98, Seattle, WA, June 1998.
- S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. M. Kleinberg. Mining the web's link structure. COMPUTER, 32:60-67, 1999.

References (2)

- J. Chen, D. DeWitt, F. Tian, and Y. Wang. NiagraCQ: A scalable continuous query system for internet databases. SIGMOD'00, Dallas, TX, May 2000.
- C. Chatfield. The Analysis of Time Series: An Introduction, 3rd ed. Chapman and Hall, 1984.
- S. Chakrabarti. Data mining for hypertex: A tutorial survey. SIGKDD Explorations, 1:1-11, 2000.
- S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. J. American Society for Information Science, 41:391-407, 1990.
- M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander. Algorithms for claracterization and trend detection in spatial databases. KDD'98, New York, NY, Aug. 1998.
- M.J. Egenhofer. Spatial Query Languages. UMI Research Press, University of Maine, Portland, Maine, 1989.
- M. Ester, H.-P. Kriegel, and J. Sander. Spatial data mining: A database approach. SSD'97, Berlin, Germany, July 1997.
- C. Faloutsos. Access methods for text. ACM Comput. Surv., 17:49-74, 1985.
- U. M. Fayyad, S. G. Djorgovski, and N. Weir. Automating the analysis and cataloging of sky surveys. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.
- R. Feldman and H. Hirsh. Finding associations in collectionds of text. In R. S. Michalski, I. Bratko, and M. Kubat, editors, "Machine Learning and Data Mining: Methods and Applications", John Wiley Sons, 1998.

References (3)

- C. Faloutsos and K.-I. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. SIGMOD'95, San Jose, CA, May 1995.
- D. Florescu, A. Y. Levy, and A. O. Mendelzon. Database techniques for the world-wide web: A survey. SIGMOD Record, 27:59-74, 1998.
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
- C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. SIGMOD'94, Minneapolis, Minnesota, May 1994.
- M. Flickner, H. Sawhney, W. Niblack, J. Ashley, B. Dom, Q. Huang, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, S. Steele, and P. Yanker. Query by image and video content: The QBIC system. IEEE Computer, 28:23-32, 1995.
- S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. ICDE'99, Sydney, Australia, Mar. 1999.
- R. H. Gueting. An introduction to spatial database systems. The VLDB Journal, 3:357-400, 1994.
- J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. ICDE'99, Sydney, Australia, Apr. 1999.
- J. Han, K. Koperski, and N. Stefanovic. GeoMiner: A system prototype for spatial data mining. SIGMOD'97, Tucson, Arizona, May 1997.

References (4)

- J. Han, S. Nishio, H. Kawano, and W. Wang. Generalization-based data mining in object-oriented databases using an object-cube model. Data and Knowledge Engineering, 25:55-97, 1998.
- J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu. Freespan: Frequent pattern-projected sequential pattern mining. KDD'00, Boston, MA, Aug. 2000.
- J. Han, N. Stefanovic, and K. Koperski. Selective materialization: An efficient method for spatial data cube construction. PAKDD'98. Melbourne, Australia, Apr. 1998.
- J. Han, Q. Yang, and E. Kim. Plan mining by divide-and-conquer. DMKD'99, Philadelphia, PA, May 1999.
- K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. SSD'95, Portland, Maine, Aug. 1995.
- J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of ACM, 46:604-632, 1999.
- E. Knorr and R. Ng. Finding aggregate proximity relationships and commonalities in spatial data mining. IEEE Trans. Knowledge and Data Engineering, 8:884-897, 1996.
- J. M. Kleinberg and A. Tomkins. Application of linear algebra in information retrieval and hypertext analysis. PODS'99. Philadelphia, PA, May 1999.
- H. Lu, J. Han, and L. Feng. Stock movement and n-dimensional inter-transaction association rules. DMKD'98, Seattle, WA, June 1998.

References (5)

- W. Lu, J. Han, and B. C. Ooi. Knowledge discovery in large spatial databases. In Proc. Far East Workshop Geographic Information Systems, Singapore, June 1993.
- D. J. Maguire, M. Goodchild, and D. W. Rhind. Geographical Information Systems: Principles and Applications. Longman, London, 1992.
- H. Miller and J. Han. Geographic Data Mining and Knowledge Discovery. Taylor and Francis, 2000.
- A. O. Mendelzon, G. A. Mihaila, and T. Milo. Querying the world-wide web. Int. Journal of Digital Libraries, 1:54-67, 1997.
- H. Mannila, H Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery, 1:259-289, 1997.
- A. Natsev, R. Rastogi, and K. Shim. Walrus: A similarity retrieval algorithm for image databases. SIGMOD'99, Philadelphia, PA, June 1999.
- B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. ICDE'98, Orlando, FL, Feb. 1998.
- M. Perkowitz and O. Etzioni. Adaptive web sites: Conceptual cluster mining. IJCAI'99, Stockholm, Sweden, 1999.
- P. Raghavan. Information retrieval algorithms: A survey. In Proc. 1997 ACM-SIAM Symp. Discrete Algorithms, New Orleans, Louisiana, 1997.

References (6)

- D. Rafiei and A. Mendelzon. Similarity-based queries for time series data. SIGMOD'97, Tucson, Arizona, May 1997.
- G. Salton. Automatic Text Processing. Addison-Wesley, 1989.
- J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, 1:12-23, 2000.
- P. Stolorz and C. Dean. Quakefinder: A scalable data mining system for detecting earthquakes from space. KDD'96, Portland, Oregon, Aug. 1996.
- G. Salton and M. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
- V. S. Subrahmanian. Principles of Multimedia Database Systems. Morgan Kaufmann, 1998.
- C. J. van Rijsbergen. Information Retrieval. Butterworth, 1990.
- K. Wang, S. Zhou, and S. C. Liew. Building hierarchical classifiers using class proximity. VLDB'99, Edinburgh, UK, Sept. 1999.
- B.-K. Yi, H. V. Jagadish, and C. Faloutsos. Efficient retrieval of similar time sequences under time warping. ICDE'98, Orlando, FL, Feb. 1998.
- C. T. Yu and W. Meng. Principles of Database Query Processing for Advanced Applications. Morgan Kaufmann, 1997.

References (7)

- B.-K. Yi, N. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, and A. Biliris. Online data mining for co-evolving time sequences. ICDE'00, San Diego, CA, Feb. 2000.
- C. Zaniolo, S. Ceri, C. Faloutsos, R. T. Snodgrass, C. S. Subrahmanian, and R. Zicari. Advanced Database Systems. Morgan Kaufmann, 1997.
- O. R. Za"iane and J. Han. Resource and knowledge discovery in global information systems: A preliminary design and experiment. KDD'95, Montreal, Canada, Aug. 1995.
- O. R. Za"iane and J. Han. WebML : Querying the world-wide web for resources and knowledge. WIDM'98, Bethesda, Maryland, Nov. 1998.
- O. R. Za"iane, J. Han, Z. N. Li, J. Y. Chiang, and S. Chee. MultiMedia-Miner: A system prototype for multimedia data mining. SIGMOD'98, Seattle, WA, June 1998.
- O. R. Za"iane, J. Han, and H. Zhu. Mining recurrent items in multimedia with progressive resolution refinement. ICDE'00, San Diego, CA, Feb. 2000.
- M. J. Zaki, N. Lesh, and M. Ogihara. PLANMINE: Sequence mining for plan failures. KDD'98, New York, NY, Aug. 1998.
- X. Zhou, D. Truffet, and J. Han. Efficient polygon amalgamation methods for spatial OLAP and spatial data mining. SSD'99. Hong Kong, July 1999.
- O. R. Za"iane, M. Xin, and J. Han. Discovering Webaccess patterns and trends by applying OLAP and data mining technology on Web logs. ADL'98, Santa Barbara, CA, Apr. 1998.