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

May 21, 2003

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 documents

May 21, 2003

Mining Complex Types of Data

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

May 21, 2003

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

May 21, 2003

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

May 21, 2003

Basic Measures for Text Retrieval

- Precision: the percentage of retrieved documents that are in fact relevant to the query (i.e., “correct” responses)
  \[
  \text{precision} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Retrieved}|}
  \]
- Recall: the percentage of documents that are relevant to the query and were, in fact, retrieved
  \[
  \text{recall} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Relevant}|}
  \]

May 21, 2003
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.

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 $\text{freq}_{ij} = \#$ of occurrences of the word $i$ 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:
    $\text{sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$

Document term matrix: example

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

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

M = 10x6 matrix

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) technique to reduce the size of frequency table.
- Retain the $K$ most significant rows of the frequency table.
- Index creation: Store the set of all vectors, indexed by one of a number of techniques (such as TV-tree).
LSI

- \( M = U S V^T \)
- \( U = 10 \times 6 \) matrix where each row is a vector of weights for a particular document
- \( S = 6 \times 6 \) diagonal matrix of eigenvalues for each principal component direction
- \( V^T = 6 \times 6 \) 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

<table>
<thead>
<tr>
<th>( D_1 )</th>
<th>( D_2 )</th>
<th>( D_3 )</th>
<th>( D_4 )</th>
<th>( D_5 )</th>
<th>( D_6 )</th>
<th>( D_7 )</th>
<th>( D_8 )</th>
<th>( D_9 )</th>
<th>( D_{10} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>30.9</td>
<td>-11.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.3</td>
<td>-10.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-7.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.4</td>
<td>-3.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52.7</td>
<td>-20.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.2</td>
<td>21.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.8</td>
<td>21.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.5</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.5</td>
<td>17.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19.9</td>
<td>45.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These are the two directions having the maximum variance of the data
Note that \( d_1 \) (database) and \( d_2 \) (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
- Use 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

MPEG-7 standard for Image/Video Representation

- MPEG-1: Storage of moving picture and audio on storage media (CD-ROM) 11 / 1992
- 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
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

Description independent of the data support

Existing solutions for textual content or description

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, 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

Scope of MPEG-7

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

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

MPEG-7 working areas

Visual Descriptors
- Color
- Texture
- Shape
- Motion

1. Histogram
   - Scalable Color
   - Color Structure
   - GOF/GOP
2. Dominant Color
3. Color Layout

- Texture Browsing
- Homogeneous texture
- Edge Histogram
- Contour Shape
- Region Shape
- Camera motion
- Motion Trajectory
- Parametric motion
- Motion Activity
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 \( \text{Rank}(k) \) 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:
\[
AVR(q) = \frac{\sum_{k=1}^{NG(q)} \text{Rank}(k)}{NG(q)}
\]

Compute the modified retrieval rank as follows:
\[
MRR(q) = AVR(q) - 0.5 \cdot \frac{NG(q)}{2}
\]
Normalized MRR, \( \text{NMRR} = MRR(q) / \text{Norm}(q) \)
Where \( \text{Norm}(q) = 1.25 * K - 0.5 - 0.5 * NG(q) \)
\[
ANMRR = \frac{1}{Q} \sum_{q=1}^{Q} \text{NMRR}(q)
\]

Texture based similarity search

Web image search

Applications - Web Image Search

Shape Descriptors

Contour-based shape descriptor
Region-based shape descriptor
Experimental Dataset & Procedure

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

Retrieval Example

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 this category

Relevance Feedback Implementation

- Weight Matrix update
  - User’s feedback: $W_0 \rightarrow W_1$
  - NN search
    - Similarity measure: $d(Q,F,W) = (Q-F)^TW(Q-F)$
  - Efficient, linear mapping
  - Kernel-based learning
    - non-linear mapping $\phi(F)$
    - NN search in the mapped space: $F \rightarrow \phi(F)$
  - Similarity measure: $d(\phi(Q),\phi(F)) = (\phi(Q)-\phi(F))^T(\phi(Q)-\phi(F))$
  - Computationally expensive, more effective

Weight Matrix Update

Weighted Euclidean Distance

- $K'$ - number of identified relevant objects
- $X = [x_1, x_2, ..., x_M]$ - relevant vectors
- $\sigma_m$ - standard deviation of the sequence of
- Diagonal weight matrix update (MARS):
  \[(W'_i)_m = \left(\prod_{k=1}^{M} \sigma_m^2\right)^{-1/2}\]
  \[\left(\prod_{k=1}^{M} \sigma_m^2\right)^{-1/2}\]
- Distance measure:
  \[d(Q,F,W) = (Q-F)^TW(Q-F)\]

Weight Matrix Update

Quadratic Distance

- Weight matrix: $W = P^T \lambda P$
  - real, symmetric, positive definite
- Mahalanobis distance update (MindReader):
  - Sample covariance matrix:
    \[C = P_m^T \lambda_m P_c\]
  - \[C_{ij} = \sum_{k=1}^{M} \pi_k (x_i - q_k)(x_j - q_k) \]
  - Weight matrix update:
    \[W = (\det(C))^{-1/2} C^{-1} \rightarrow P = P_c \lambda_m \sum_{k=1}^{M} \pi_k \lambda_k\]
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

- Will make the paper available online

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: \( N=\{A, D, E, F, G, H\} \)
  - Phase 2: \( R=\{G, H\} \)
- Objective: minimize set of candidates that contains all the \( K \) nearest neighbors

K-NN Search in VA-File

- Bounds: \( L(Q,W) \leq d(Q,F,W) \leq U(Q,W) \)
- 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 (\( N_1 \) candidates)
  - Phase II filtering (data level): visit the \( N_1 \) 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 (\( N_2 \) feature vectors)

Bound Computation in Relevance Feedback

- Quadratic distance
- Weighted Euclidean distance
Bound Computation

Quadratic Distance

- Distance: $d(Q, F, W) = (Q - F)^T W (Q - F)$
- Weight Matrix: $W = P^T \lambda P$
- Mapping: $D^i = PD$
- 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
  - Filter at 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: $R_t = \{ F^i_j, j \in [1, K] \}$
- Upper bound of that set: $r_t(Q) = \max \{ d(Q, F^i_j, W) \}$
  - If $t>1$, define: $r^*_t(Q) = \max \{ d(Q, F^i_{j-1}, W) \}$
  - Then: $r_t(Q) \leq r^*_t(Q)$
- Maximum of $K$ distances between the query $Q$ and objects in $R_t$ computed using $W$ can not be larger than $r^*_t(Q)$
  - $L_t(Q, W) \leq d(Q, F, W) \rightarrow \max \{ L_t^{i_1}(Q, W) \} \leq r^*_t(Q)$

A tighter bound on NN set

- In similar manner, for $t>1$, define $L_t^i(Q) = \max \{ L_t^i(Q, W) \}$
- Then: $\max(L_t^{i_1}(Q, W)) \leq L_t^i(Q)$
- Maximum of the lower bounds between the query $Q$ and objects in $R_t$ computed using $W$ can not be larger than $L_t^i(Q)$
- Therefore: $L_t(Q, W) \leq d(Q, F, W) \rightarrow L_t^i(Q) \leq r^*_t(Q)$
- Filtering bounds: $L_t(Q, W) \leq L_t^i(Q) \leq r^*_t(Q)$

Adaptive NN Search Phase I filtering

1) if $t=1$, use the traditional NN search; BREAK;
2) Given $R_t$ and $W_t$ compute $r^*_t(Q)$ and $L_t^i(Q)$
3) for $i=1$ to $N$
   a) Compute $L_t(Q, W_t)$ for weighted distance
   b) Compute $L_t^i(Q, W_t) = L_t^i(Q, W_t)$ for quadratic dist.
      - If $L_t(Q, W_t) \leq r^*_t(Q)$ insert $C(F)$ into $N^{l_i}_t(Q, W_t)$
      - If $L_t(Q, W_t) \leq L_t^i(Q)$ insert $C(F)$ into $N^{l_i}_t(Q, W_t)$

Experiments

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

- 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. PKDD'00, Melbourne, Australia, Apr. 1999.
- J. Han, Q. Yang, and E. Kim. Plan mining by divide-and-conquer. PKDD'00, Philadelphia, PA, May 1999.

References (3)

- J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series databases. ICDE'98, Sydney, Australia, Apr. 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. PKDD'00, Melbourne, Australia, Apr. 1998.
- J. Han, Q. Yang, and E. Kim. Plan mining by divide-and-conquer. PKDD'00, Philadelphia, PA, May 1999.
- 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. PKDD'00, Melbourne, Australia, Apr. 1998.
- J. Han, Q. Yang, and E. Kim. Plan mining by divide-and-conquer. PKDD'00, Philadelphia, PA, May 1999.
References (5)


References (6)


References (7)