Chapter 21: New Applications

- The World Wide Web
- Distributed Information Systems
- Information-Retrieval Systems
- Mobility and Personal Databases
- Multimedia Databases
- Spatial and Geospatial Databases
- Data Warehousing
- Data Mining
- Data Analytics
- Decision-Support Systems
A data warehouse archives information gathered from multiple
sources, and stores it under a unified schema, at a single site.

By using data mining, we can discover knowledge automatically in the
database.

Can interface statistical analysis packages (S+) with
SQL queries, and by SQL extensions.

Information for decision support can be extracted by simple
transaction-processing systems to generate summary data.

Decision-support systems utilize data collected by on-line
Can represent the data in relational form by using the value all to represent subtotals.

<table>
<thead>
<tr>
<th></th>
<th>88</th>
<th>15</th>
<th>45</th>
<th>28</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dark</td>
<td>35</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Light</td>
<td>53</td>
<td>10</td>
<td>35</td>
<td>8</td>
<td>90</td>
</tr>
<tr>
<td>Small</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>16</td>
<td>51</td>
<td>38</td>
<td>103</td>
</tr>
</tbody>
</table>

Sales with the schema, Sales (color, size, number).

Cross-tabulation of number by size and color of sample relation ranges, and computes an aggregate over the values in each range, and computes an aggregate over the values taken by an attribute into a histogram partitions the values taken by an attribute into.

Support simple data analysis.

Aggregate functions summarize large volumes of data to

Data Analysis
the value of number by sum.
individually tuples with different values for size and by replacing

obtain (light, ali, 53) and (dark, ali, 35) by eliminating

<table>
<thead>
<tr>
<th>Number</th>
<th>Size</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>ali</td>
<td>ali</td>
</tr>
<tr>
<td>15</td>
<td>large</td>
<td>ali</td>
</tr>
<tr>
<td>45</td>
<td>medium</td>
<td>ali</td>
</tr>
<tr>
<td>28</td>
<td>small</td>
<td>ali</td>
</tr>
<tr>
<td>35</td>
<td>ali</td>
<td>dark</td>
</tr>
<tr>
<td>5</td>
<td>large</td>
<td>dark</td>
</tr>
<tr>
<td>10</td>
<td>medium</td>
<td>dark</td>
</tr>
<tr>
<td>20</td>
<td>small</td>
<td>dark</td>
</tr>
<tr>
<td>33</td>
<td>ali</td>
<td>light</td>
</tr>
<tr>
<td>10</td>
<td>large</td>
<td>light</td>
</tr>
<tr>
<td>35</td>
<td>medium</td>
<td>light</td>
</tr>
<tr>
<td>8</td>
<td>small</td>
<td>light</td>
</tr>
</tbody>
</table>

Data Analysis (Cont.)
GROUP BY color, size, with cube
FROM sales
SELECT color, size, sum(number)

The following query generates the previous table.

To support generation of summary data, proposed extensions to SQL, such as the cube operation, help

- Inner-granularity data.
- Drill down: Moving from coarse-granularity data to
- Coarser granularity by means of aggregation.
- Rollup: Moving from inner-granularity data to a coarser
2. The system is responsible for automatically discovering knowledge from the database, by detecting patterns and correlations in the data.

1. The user is involved directly in the process of knowledge discovery.

Discover rules using one of two models:

- A set of rules.
- Knowledge discovered from a database can be represented by a primarily on disk.

Learning in that it deals with large volumes of data, stores discoveries statistically rules and patterns; it differs from machine learning. Unlike knowledge discovery in artificial intelligence, data mining...
Confidence: Measure of how often the consequent is true when both the antecedent and the consequent of the rule.

Support: Measure of what fraction of the population satisfies the rule.

Population: Cross-product of the ranges of the variables in the transaction database.

The rule \( \text{transactions} \backslash \text{buys}(\text{bread}, \text{milk}) \iff \text{buys}(\text{bread}) \backslash \text{buys}(\text{milk}) \)

\( X \) is a list of one or more variables with associated ranges.

The general form of rules: \( X \) and consequent \( \iff \) antecedent
go up, stock prices go down within two days.
information and related sequenced data (e.g. When bond rates
Determination: determine correlations between
someone who buys bread is quite likely also to buy milk).
Useful to determine associations between different items (e.g.,
worthiness).
which of several factors help classify a person’s credit
distinct groups that are relevant for making a decision (e.g.,
Classification: Finding rules that partition the given data into
Classes of Data-Mining Problems
Data-visualisation detects patterns in large volumes of data via

C. Credit = Excellent
000 P. Degree = Master's and C. Income ≥ 75,000

A people P, P. degree = Master's and C. income ≥ 75,000

Example: Reason the hypotheses: "People who hold master's degrees are the most likely to have an excellent credit rating."

Rule expressing the hypotheses:

Derive from the database the confidence and support for the hypotheses.

Primary responsibility for discovering rules is with the user.

User-Guided Data Mining
In general, different branches of the tree could grow to different levels. Different branches stop here for each partition based on degree. 

The construction stops here by income adequately classify the tuples. 

Credit example: Within each partition based on degree, the attributes have been considered. 

The construction of a path in the tree stops when either the target attribute has properly classified the data, or all attributes on one attribute. 

Each node of the tree partitions the data into groups based on the target attribute. 

Top down generation of classification tree. 

Tuple is already known. 

Training set: a data sample in which the grouping for each
\(A \text{ transactions, } \{L\}\text{shinq } \leq \cup \text{ , \(L\}\text{shinq } \text{and } \cdots \text{ and } \{L\}\text{shinq} \text{ transactions.}
\)

Consider all subsets of the set of relevant items, and for each

\(\{L\}\text{shinq } \leq \cup \text{ , \(L\}\text{shinq } \text{ and } \cdots \text{ and } \{L\}\text{shinq} \text{ transactions.}
\)

Usually, these rules will involve a single item of interest in the shop.

\(\{L\}\text{shinq } \text{ and } \{L\}\text{shinq} \text{ transactions.}
\)

Derive rule by associating a bitmap with each transaction with

\(\{L\}\text{shinq } \leq \cup \text{ , \(L\}\text{shinq} \text{ bread } \cup \text{ , \(L\}\text{shinq} \text{ milk} \}
\)

Discovery of Association Rules
be considered.

fraction of the transactions, none of its supersets needs to
- Once a set is eliminated because it occurs in too small a
- Use multiple passes, considering only some sets in each pass.

corresponding to large.

Many sets: Cost of processing each transaction becomes
- Items with a high degree of association to
- Sets with a high count at the end of the pass correspond to
- Transaction's bitmap.

each set of items, all of whose bits are set in the
- When a transaction is fetched, the count is incremented for
- A count is maintained for each set, initially set to 0.

Few sets: Determine level of support via a single pass.

Discovery of Association Rules (Cont.)
What data to summarize.

– How to propagate updates.

– What schema to use.

– When and how to gather data.

Issues in building a warehouse:

• decision-support workload.

Decision-support systems are not affected by the transaction-processing systems that on-line warehouse, the decision maker ensures that on-line warehouse, the decision support for decision support from a data.

• By accessing information for decision support easier to write.

• Provides a single consolidated interface to data, making
• Geographic databases store geographic information (e.g., maps) or, often called geographic information systems or GIS.

• Geographic databases or layouts of integrated circuits.

• Constructed or layouts of objects (e.g., buildings, art). are information about how objects (e.g., buildings, art) are

• Design databases (also CAD) databases store design

• Spatial data, and for processing spatial join queries.

• Spatial data are important for accessing

• Special purpose index structures are important for accessing

• Spatia to spatial databases store information related to spatial locations

• Spatia and Geographic Databases
Alternative: Triangulation — give polygons an identifier; each
same as the ending vertex.
Closed polygons: List its vertices in order, starting vertex is the
carries with it the identifier of the curve (2D features such as
segments, represent each segment as a separate tuple that also
Approximate a curve by partitioning it into a sequence of
Represent a line segment by the coordinates of its endpoints.
Various geometric constructs can be represented in a database.
polypolygon.

with an indication of which side of the face is inside the

Alternative: List their faces, each of which is a polygon, along
tetrahedrons, like triangulating polygons.

Represent arbitrary polyhedra by dividing them into

2-D, except that points have an extra $z$ component.

Representation of points and line segments in 3-D similar to

Representation of Geometric Information (Cont.)
representation

\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}
\{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), (x_5, y_5)\}
\{(x_1, y_1), (x_3, y_3), (x_4, y_4), ID_1\}
\{(x_1, y_1), (x_4, y_4), (x_5, y_5), ID_1\}

object

polygon

\{(x_1, y_1), (x_3, y_3), (x_2, y_2)\}

polygon

\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}

triangle

\{(x_1, y_1)\}

line segment

\{(x_1, y_1), (x_2, y_2)\}

Representation of Geometric Constructs
of simpler objects.

Viewport models represent three-dimensional surfaces as a set

Intersection, and difference operations,
objects such as spheres, cylinders, and cuboids, by union,

Complex three-dimensional objects: formed from simpler

via union, intersection, and difference operations.

Complex two-dimensional objects: formed from simpler objects

rectangles, polygons.

Simple two-dimensional objects: points, lines, triangles,

design is structured.

Represent design components as objects (generally geometric)

Design Databases
Design databases generally do not store raster data.

- Taken at different points in time.
- Different altitudes at different regions, or measurements at different dimensions might include the temperature at area.
- Example 2-D raster image: satellite image of cloud cover,
- Image dimensions.

- Raster data consist of bit maps or pixel maps, in two or more Geographic Data.
- Features such as regions and lakes can be depicted as polygons.
- Roads can be considered as two-dimensional and represented by lines and curves.
- Some features, such as rivers, may be represented either as complex curves or as complex polygons, depending on whether their width is relevant.

Vector data are constructed from basic geometric objects:

- Vector format often used to represent map data.
- Vector data are constructed from basic geometric objects: points, line segments, triangles, and other polygons in two dimensions, and cylinders, spheres, cuboids, and other polyhedrons in three dimensions.
an accuracy of tens of meters. Broadcast from GPS satellites to find the current location with GPS unit – utilizes information.

- Global Positioning System (GPS)
- Vehicle navigation systems store information about roads and
  - Geographic Information Systems (GIS), provide location.

Geographic Data (Cont.)
Spatial Queries

- attribute
- two spatial relations with the location playing the role of join
- queries that compute intersections of regions - spatial join of queries that compute intersections of regions - spatial join of queries that compute intersections of regions - spatial join of queries that compute intersections of regions - spatial join of queries that compute intersections of regions - spatial join of
- that lie partially or fully inside a specified region
- Region queries deal with spatial regions, e.g., ask for objects
- Region queries deal with spatial regions, e.g., ask for objects
- Nearest queries request objects that lie near a specified
- Nearest queries request objects that lie near a specified
Spatial Queries (cont.)

• Spatial data is typically queried using a graphical query language, results are also displayed in a graphical manner.

   – Spatial data can mix spatial and non-spatial conditions
   – Graphical interface constitutes the front-end; extensions of SQL with abstract data types, such as lines, polygons and bit maps

   – Supports can mix spatial and non-spatial conditions
   – Extensions also include and allowing spatial conditions

(contains or overlaps)
nodes for each internal node; well-suited for secondary storage.

The $k$-d tree extends the $k$-tree to allow multiple child
number of points.

Partitioning stops when a node has less than a given maximum
sub-tree fall on one side and half on the other.

In each node, approximately half of the points stored in the
cycling through the dimensions.

otherwise dimension in nodes at the next level and so on,
one dimension at the node at the top level of the tree.

Each level of a $k$-d tree partitions the space into two.
dimensions.

$k$-d tree — early structure used for indexing in multiple

Indexing of Spatial Data
The tree at which the corresponding node appears.

The number of the lines in the figure indicates the level of correspondence to a node in the k-d tree.

Each line in the figure (other than the outside box)

Each line corresponds to a node in the tree, and the maximum number of points in a leaf node has been set at 1.
number of points (set to \( 1 \) in example).

- Leaf nodes have between zero and some fixed maximum nodes corresponding to the four quadrants.
- Each non-leaf node divides its region into four equal sized quadrants, and corresponds to each such node has four children.
- Each node is associated with a rectangular region of space; the quadtree.
segments and polygons.

Extrusions of K-d trees and quadtrees handle indexing of line
which have the same value.

single array element, or have multiple array elements, all of
- The subarrays corresponding to leaves either contain just a
  each node corresponds to a subarray of values.
  internal node.
  further into four children of equal area, and is therefore an
  that if covers are the same. Otherwise, it is subdivided
  a node is a leaf node if all the array values in the region
  Region quadtrees store array (raster) information.
  rather than on the actual set of points stored
  quadtree stores points' space is divided based on regions,

Quadtree (cont.)
or quadruples since a polygon is stored only once.

- The storage efficiency of R-trees is better than that of K-d trees.
- A polygon must contain the polygon.
- A polygon is stored only in one node, and the bounding box of
  the node must contain the polygon.
- A bounding box associated with all its children.
- The bounding box associated with a non-leaf node contains the
  bounding box associated with each tree node.
- A rectangular bounding box is associated with each tree node.
- A rectangular and other polygons.
- Variants like R+-trees and R*-trees are useful for indexing.
- R-trees are a two-dimensional extension of B-trees and, with
The R-Tree is shown on the right. Line (solid line) and the bounding boxes (dashed line) of the nodes of an R-Tree for the set of rectangles. A set of rectangles (solid line) and the bounding boxes (dashed line).
Continuous media data

Must provide guaranteed steady retrieval rates for

Structures

Similarity-based retrieval must be provided by special index

The database must handle large object representation

then storing them outside the database, in a file system

It is desirable to store multimedia data in a database (rather

To provide such database functions as indexing and consistency
Handwritten data — Identity a handwritten data item or

Audio data — Speech-based user interfaces allow the user to

Pictorial data — Two pictures or images that are slightly

Similarity-Based Retrieval
moving synchronously with the audio).

maintained (video of a person speaking must show lips

– Synchronization among distinct data streams must be

overload of system buffers.

– Data must be delivered at a rate that does not cause

audio or video result.

– Data must be delivered sufficiently fast that no gaps in the


Information-delivery requirements:

characterized by high data volumes and real-time

• Most important types are video and audio data.

Continuous-Media Data
Compresses 1 minute of audio-video to approximately 2.25 MB.

- Video disks' negligible loss of video quality.
- MECE-2 designed for digital broadcast systems and digital video using only JPEG.
- Quality comparable to VHS video tape.
- MECE-1 stores a minute of 30-frame-per-second video and audio in approximately 1.25 MB (compares with 75 MB for
  frames to achieve a greater degree of compression.
- MECE standards use commonalities among a sequence of
  successive frames of a video are often nearly the same.
- Encoding each frame of a video using JPEG is wasteful, since
  is the most widely used format for image data.

Multimedia Data Formats
Video servers

Network — Transmission of multimedia data from a server to boxes.

Head-end terminals — used to view multimedia data; PCs or access data.

Video server — Multimedia data are stored on several disks requiring high-capacity networks.

Multiple head-end terminals require a high-capacity network;
 Query result: user's locations may be a parameter of the query. Difficulty to determine the optimal location to materialize a

Without fixed locations and network addresses, can become

Handoff of control from one mobile support station to another

Mobile hosts may move between cells, thus necessitating a

Within its cell

Each mobile support station manages those mobile hosts

Referred to as mobile support stations

Mobile hosts communicate to the wired network via computers

Referred to as mobile hosts, and a wired network of

Computers

The mobile computing environment consists of mobile

Mobility and Personal Databases
Energy to receive than to transmit radio signals.

Data transfer present different power demands (requires less power per bit in digital cellular systems.)

Energy — optimize use of battery power; different forms of power.

Periods

Time-of-day based charges — vary based on peak or off-peak charges in digital cellular systems.

Number of bytes, or packets, transferred — used to compute connection time — used to assign monetary charges in some systems.

User time.

Must consider these computing costs:
well-known time intervals.

be broadcast at a well-known radio frequency and at

— For changeable schedule - the broadcast schedule must itself

schedule or a changeable schedule.

Broadcast data may be transmitted according to a fixed

— Relevant data will be broadcast.

— Transmit a request for data and must know when the

— Wait for the data to be broadcast

must either:

A query can be answered using only cached data; if not then

— A mobile host may optimize energy costs by determining if a

having to consume energy transmitting a request,

data; allows mobile hosts to wait for needed data, rather than

Mobile support stations can broadcast frequently requested
the mobile host cannot discover this until it is reconnected.

**Consistency:** Cached data may become out of date, but storage cannot be simulated well.

**Rercoverability:** Updates entered on a disconnected mobile host may be lost if the mobile host fails. Since the machine may be lost if the mobile host fails. Since the

and updates on data that resides or is cached locally:

- Problems created if the user of the mobile host issues queries disconnected.

- A mobile host may remain in operation during periods of

**Disconnection and Consistency**
Facilitating the reconciliation of inconsistent copies of data still under research.

Information either directly or through a common host. The clash will be detected eventually, when the hosts exchange hosts independently. Update the same version of a document.

Version-numbering-based schemes guarantee only that if two entities, however, mobile host may miss a report.

Reports inform a reconciled mobile host of out-of-date cache when data are updated by other computers, invalidation becomes invalid and updates may conflict.

Update when mobile host reconciles; in other cases data may be propagated.

For data updated by only the mobile host, simple to propagate in mobile computing.

Partitioning via disconnection is the normal mode of operation.

Mobile Updates
database).

Similarity-based retrieval – similarity may be defined based on answers based on their potential relevance.

Precisely, and hence information retrieval systems order.

The query a user has in mind usually cannot be stated very

specifically, for example, sets of keywords.

Queries attempt to locate documents that are of interest by

capabilities within the restricted model.

Information retrieval systems use a simpler data model than

Information Retrieval Systems
query were retrieved.

Recall — what percentage of the documents relevant to the query are relevant to the query.

Precision — what percentage of the retrieved documents are relevant.

Relevant performance metrics: •

Few false positives.

Index should not permit any false drops, but may permit a few irrelevant documents may be retrieved.

false positive — a few irrelevant documents may be retrieved.

false drop — a few relevant documents may not be retrieved.

Storing the index for approximate retrieved saves space. •

Set $S$, of documents that contain $Y$.

set $S$, of documents that contain $Y$, which maps each keyword $Y$, to the inverted index, which maps each keyword $Y$ to the documents that contain a specified keyword can be located.

Indexing of Documents
of documents that contain the keyword $K$.

by taking the difference $S - S'$, where $S$ is the set of identifiers specified keywords $K$. Given a set of document identifiers $S$, we can eliminate documents that contain the specified keyword $K$.

but not operation — finds documents that do not contain a

operation: $S = S \cup S' \cup \cdots \cup S_n$.

union, $S = S_1 \cup S_2 \cup \cdots \cup S_n$, of the sets.

by computing the least one of the keywords $K_1, K_2, \cdots, K_n$ gives the set of all documents that contain at

or operation — finds documents that contain any

and operation — finds documents that contain all of a set of

Indexing of Documents (Cont.)
WWW. 

Stored at other sites in a distributed system such as the World

Distributed hypertext systems permit references to documents

media such as images, videos, and audio clips.

Hypermedia systems provide not only text, but also other

retrieves and displays it.

Hypermedia systems provide a Facilities where the user can easily

Typically use a point-and-click interface, where a simple mouse

switch from one document to another.

Hypermedia systems take the idea of storing document identities

Browsing and Hypertext
GUIs. Provide standardized ways of accessing data and standardized distributed heterogeneous systems have also been developed;

- Automated tools for locating and indexing information in such information using the hypertext paradigm.

- The World Wide Web system supports browsing such

- Distributed information systems running on the Internet have seen explosive growth in recent years.
information stored at other sites, and how to access them.

Each site maintains a site description, describes the kind of

via a powerful keyword-based indexing mechanism.

Wide Area Information System (WAIS) retrieves information

navigating a directory hierarchy.

Information retrieved in Gopher is based on browsing and

- Allows a seamless connection to remote servers.
- Allows a link to a directory on another server.
- A menu item can be another directory in the hierarchy, a
- Menu item initializes a directory hierarchy with a server, the top level
- Server organizes data into directories.

A Gopher system consists of servers and clients.

The Gopher and WAIS
The World Wide Web

can be active, rather than just passive.
The programming language java allows documents to contain programs that are executed at the user’s site, thus documents

• link is then displayed.

Link associated with it, and the document pointed to by the
appropriate browser, the user can click on a region that has a

• The displayed document is a hypertext document with an

  the standard Generalized Markup Language (SGML), which is based on

Most Web documents are hypertext documents formatted via

  A distributed information system based on hypertext.
machine.

- The rest of the URL is the path name of the file on the
  Internet.
- The second part gives the unique name of a machine on the
  HyperText Transfer Protocol.
- "http" indicates that the document is to be accessed using
  "http".
- The first part indicates how the document is to be accessed:

  URL example: http://www.bell-labs.com/100/bar

Universal Resource Locators (URLs).

In the Web, the functionality of pointers is provided by

Universal Resource Locators
Home pages – contain information about users and their work.

documents on the Web, and give the URL to anyone else.
administrators, anyone connected to the Internet can create
Cipher systems are set up and controlled by system

document and a directory.

a HTML document displayed simultaneously a
directory, a HTML document displayed is simultaneously a
With Cipher, what is displayed is either a document or a
The HTML language supported by the Web provides a graphical user interface to the information service.

- To install a new service on the Web, one simply needs to create a document.
- When a HTTP server receives a request for such a document, it executes a program that generates a special executable HTML document.
- The document name in a URL may identify an executable information service.
- A Web server can easily serve as a front end to a variety of

**Web Servers**
In addition to formatting, HyperText Link, and Image Display commands, HTML provides some limited input features.

- SCML provides a grammar for specifying document formats based on standard markup annotations.
- SCML provides a grammar for specifying page description/text formatting languages.
- Markup Language (SCML) is a void between plain text and Text markup languages such as the Standard Generalized Display Languages.
the option to abort the program or to continue execution. Potentially dangerous actions, such as file writes, and allows
make any system calls directly. If nothing the user about
Java's security system ensures that the Java code does not

- Site for processing.
- Greatly, compared to every interaction being sent to a server
- Executing programs at the client site speeds up interaction
- Flexible interaction with the user.

Benefits of Java:

- Client site.
- Documents, and can be downloaded and executed by any
  documents, and can be stored at server sites (like HTML
- Java programs can be stored at the local site).
- Java language allows documents to be active (e.g., animation
- Java
multiple Web documents replicate data.

- Problematic to update Web documents, especially if

- Cannot customize fixed Web documents for individual users.

- Fixed HTML source for display to users have limitations:

  - Formatted into HTML and displayed to the user.
  - Executed as a database transaction, the results can be
  - Example: Information filled in on an HTML order form can be
  - Processing with the Web.

  - Extremely useful to link databases used for transaction

Web Interfaces to Databases
documents will be updated too. Whenever relevant data in the database are updated, the
data in the database is also updated. If necessary, the data in the database
is updated.

* Tailor the display based on user information stored in the
database.

* Generate a document based on the results.

* Execute a program to run queries on the database and to
execute a script — a document request

Web Interfaces to Databases (Cont.)
When the document is requested, a macro processor executes embedded SQL queries.

Variables defined in HTML forms can be used directly in the SQL queries.

Define a HTML document in a macro language with embedded SQL queries.

Conversions from HTML to SQL and from database results into HTML interfaces to databases simplify the tasks of format.
the documents via an index. Since no central authority is required for registering
•
discard defunct sites.
•
•
•
–
obtain updated information from known sites.
–
new sites.
–

These systems run a background process to
other documents, and build an index on the documents.

Web crawlers follow the hypertext links in documents to find

found from various sites.

The Archie system automatically follows Gopher links to locate

Locating Information on the Web.