Intro: What is datamining?

- Data are generated in large amount. E.g. transactions, telephone calls.
- Data is collected because believed to be a potential source of valuable info.
- Datamining is finding useful and interesting info from the data.
- Data can be "large" in two ways: width and height of dataset.
- At the beginning, we have the computer analyze the data and spit out result in text... Now we're moving towards "human-centred datamining," and visualization is one tool to do so.

1

- **Information Visualization and Visual Data Mining**, Keim, IEEE Transactions on Visualization and Computer Graphics 8(1), 2002.
- **DataJewel: Tightly Integrating Visualization with Temporal Data Mining**, Mihael Ankerst, David H. Jones, Anne Kao, Changzhou Wang. ICDM Workshop on Visual Data Mining, Melbourne, FL, 2003 [Archived version]
- **DEVise: Integrated Querying and Visual Exploration of Large Datasets**, Miron Livny, Raghu Ramakrishnan, Kevin Beyer, Guangshun Chen, Donko Donjerkovic, Shilpa Lawande, Jussi Myllymaki, and Kent Wenger. Proc. SIGMOD 1997.

2

Visual data mining: include the human in the data exploration process

Combines

1) the flexibility, creativity and general knowledge of the human and
2) Enormous storage capacity and computational power of computers

3

Classification of Visual Data Mining Techniques

1) Data type to be visualized (6)
2) Visualization technique (5)
3) Interaction and distortion technique (5)

These 3 dimensions of classification can be assumed orthogonal

4

---

1. Data type to be visualized (1/2)

1.1) 1-D data, usually the dimension is very dense.
   E.g. temporal data, like time series of stock prices.
1.2) 2-D data.
   E.g. geographical maps
1.3) Multi-Dimension
   E.g. tables from relational databases
   No simple mapping of attributes to the two dimensions of the screen

5

---

1. Data type to be visualized (2/2)

1.4) Text and hypertext, e.g. news articles
   Most of the standard visualization techniques cannot be applied. In most cases, a transformation of the data into description vectors is necessary first.
   E.g. word counting, then principal component analysis.
1.5) Hierarchies and graphs
   E.g. telephone calls
1.6) Algorithms and software
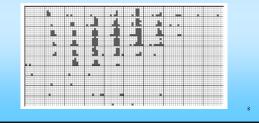   E.g. for debugging operations

6

2. Visualization technique

2.1) standard 2D/3D displays
  e.g. bar charts and x-y plots.
2.2) geometrically transformed displays
  e.g. parallel coordinates.
2.3) icon-based displays (glyphs)
2.4) dense pixel displays

7

2.5) stacked displays
  Tailored to present data partitioned in a hierarchical fashion.
  Embed one coordinate system inside another coordinate system.
  Figure: by M. Ward, Worchestor Polytechnic

8

---

3. Interaction and distortion technique (1/2)

- Dynamic: changes to visualizations are made automatically
- Interactive: changes are made manually
3.1) Dynamic projections
  e.g. To show all interesting two-dimensional projections of a multi-dimensional dataset as a series of scatter plots.
3.2) Interactive filtering
  browsing: direct selection of desired subset
  querying: specify properties of desired subsets

9

---

3. Interaction and distortion technique (2/2)

3.3) Interactive zooming
  On higher zoom levels, more details are shown.
3.4) Interactive distortion
  Show portions of the data with high level of detail while other s are shown with lower.
  E.g. spherical distortion and fisheye views.
3.5) Interactive Linking and Brushing
  – Combine different visualization methods to overcome the shortcomings of single techniques.
  – Changes to one visualization are automatically reflected in the other visualization.

10

Critiques

+ Good summary of visual datamining and InfoVis in general.
+ Nice all-around introductory material. Concise.
+ Great references. Supported his classifications with ample examples, and cites figures from other papers. "see Fig. 5 in [10]"
+ Good amount of pictures

11

- Information Visualization and Visual Data Mining, Daniel A. Keim, IEEE Transactions on Visualization and Computer Graphics 8(1), 2002.
- DataJewel: Tightly Integrating Visualization with Temporal Data Mining Mihael Ankerst, David H. Jones, Anne Kao, Changzhou Wang. ICDM Workshop on Visual Data Mining, Melbourne, FL, 2003 [Archived version]
- DEVise: Integrated Querying and Visual Exploration of Large Datasets Miron Livny, Raghu Ramakrishnan, Kevin Beyer, Guangshun Chen, Donko Donjerkovic, Shilpa Lawande, Jussi Myllymaki, and Kent Wenger. Proc. SIGMOD 1997.
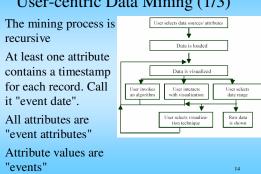
12

DataJewel

Main contribution:

- The DataJewel architecture tightly integrates a visualization component, an algorithmic component and a database component for temporal data mining.
- Bridge the field of InfoVis with other research communities e.g. datamining.
- 2 aspects of temporal data mining: Need to add new mining algorithms easily; need to link tables together that have no primary key.

13

User-centric Data Mining (1/3)

- The mining process is recursive
- At least one attribute contains a timestamp for each record. Call it "event date".
- All attributes are "event attributes"
- Attribute values are "events"

14

User-centric Data Mining (2/3)

Assumptions:

a) number of event attributes is low. (<10)

Often, in one given analysis, the analyst selects a small number of event attributes which can be associated with each other in a particular domain.

b) number of different events of one event attribute is moderate. (<200)

If this is not true, a concept of hierarchy can be defined for the event attribute.

c) smallest time unit of interest in the event dates is one day
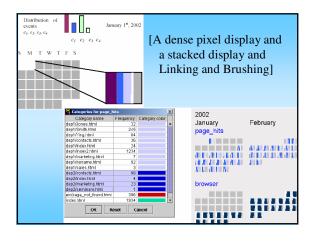
15

User-centric Data Mining (3/3)

Using the above assumptions, one instance of the visualization and the algorithmic component are presented, and new ones can be easily integrated.

16

Visualization component: CalendarView

- Multi-Dimensional, with Even Date as the "key"
- Web-mining example:

Event Date	Event Attribute: Page hit	Event Attribute: Browser	Event Attribute: ...
1/1/2002	Index.html	MS IE	...
1/1/2002	Dep1/contacts.htm	Netscape	...
...	...	...	...

17

[A dense pixel display and a stacked display and Linking and Brushing]

3

Interaction with CalendarView

- Selection: selected subset can be visualized following the iterative process
- Descending/Ascending order: good for finding "main" events and outlier events.

[Interactive filtering and interactive zooming]
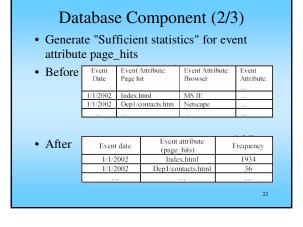
19

Temporal Mining Component

- These algorithms assign colour to events to allow users to observe patterns easily in the CalendarView.
- LongestStreak: Discover one event of one event attribute with the longest consecutive streak of significant days. (What about the longest N streaks?)
- MatchingEvents extends LongestStreak: Return the LongestStreak event and the correlated event.
- MatchingEvents2: returns the LongestStreak of the first event attribute and for each other event attribute, the event that is correlated.

20

Database Component (1/3)

- This component provide access to datasets in tables from relational database(s).
- The critical task is to scale up to large databases.
- Compute an aggregated version of the dataset such that it fits in main memory.
- Query:

```
SELECT Event_date, page_hits, count(*) as Frequency
FROM example_table
GROUP BY Event_date, page_hits
ORDER BY Event_date, page_hits;
```

21

Database Component (2/3)

- Generate "Sufficient statistics" for event attribute page_hits
- Before

Event Date	Event Attribute: Page hit	Event Attribute: Browser	Event Attribute: ...
1/1/2002	Index.html	MS IE	...
1/1/2002	Dep1/contacts.htm	Netscape	...
...	...	...	...

- After

Event date	Event attribute (page_hits)	Frequency
1/1/2002	Index.html	1934
1/1/2002	Dep1/contacts.html	36
...	...	...

22

Database Component (3/3)

- mem_init = c * number of days * average number of events per day (= 402 in aircraft maintenance domain for one airline)
- mem_new = c * number of days * average number of distinct events per day (= 32)
- Summary statistics always fit in main memory and the computation of the proposed algorithm is efficient. Authors believe it is true for most datasets which fulfill their assumptions. E.g. number of event attributes is low (<10).

23

Experiment with airplane maintenance datasets (1/2)

- Pentium III/800Mhz and 1 GB main memory
- Datasets span 12-14 years, with sufficient statistics fit in main memory
1) LongestStreak finds a system of an airplane: "engine fuel". During the last five days of July 2000, we perceive many events, indicating problems with engine fuel.

24

Experiment with airplane maintenance datasets (2/2)

2) Add several datasets to compare this finding. Manually colour every system except engine fuel with one light colour and a dark colour to all engine fuel related events: Pattern is not present.

3) Run MatchingEvents2 to single out one airplane, which has a lot of maintenance events ion Dec 3rd, 1997

4) Finally, select a dataset with maintenance events of just this plane. MatchingEvents algorithm finds fuel and communications events frequently co-occur. E.g. on Monday 18th, Nov.

5) Drill down to the raw data to further investigate. 25

Concluding remark

- Author believes the DataJewel architecture is also well adapted to areas like homeland security, market basket analysis, or intrusion detection.

26

Critique

+ Good example domains with which the DataJewel system is useful

+ Step-by-step procedure of a datamining session on airline maintenance example

- How really useful is an architecture? To use DataJewel on other domains, still need to provide algorithm, visualization (and of course dataset).

- Somewhat strong assumptions

+ The proposed algorithms can finish within 1 second -- this is over 10 years of airline maintenance data. Not bad.

- But the run time for the system as a whole -- making the sufficient statistics table and rendering is not discussed. 27

- **Information Visualization and Visual Data Mining**, Daniel A. Keim, IEEE Transactions on Visualization and Computer Graphics 8(1), 2002.
- **DataJewel: Tightly Integrating Visualization with Temporal Data Mining**, Mihael Ankerst, David H. Jones, Anne Kao, Changzhou Wang. ICDM Workshop on Visual Data Mining, Melbourne, FL, 2003 [Archived version]
- **DEVise: Integrated Querying and Visual Exploration of Large Datasets**, Miron Livny, Raghu Ramakrishnan, Kevin Beyer, Guangshun Chen, Donko Donjerkovic, Shilpa Lawande, Jussi Myllymaki, and Kent Wenger. Proc. SIGMOD 1997.

28

DEVise

- DEVise is a data exploration system that allows users to easily develop, browse, and share visual presentations of large tabular datasets from several sources.

  [Multi-dimensional datasets]

- The framework has been already successfully applied to a variety of real applications.

29

Main contributions (1/2)

1) Visual Presentation Capabilities remarkable variety to be developed easily through a point-and-click or easy-to-write 'plugins'

2) Ability to handle large (bigger than main memory), distributed (e.g. over the Web) dataset by using a declarative approach to define their visualization primitives, instead of a programming-oriented style.

3) Collaborative data analysis: several users can share visual presentations of the data and dynamically explore these presentations. 30

Main contributions (2/2)

- Visual querying from a variety of local and remote sources. From the visual representations being used, the system can dynamically gather hints for what to index, materialize, cache or re-compute.

31

Examples

- Financial data exploration in the UW Business school: look for correlations and trends using the combined information from a variety of vendors.
- R-tree validation: discover subtle bugs in the R-tree bulk loading algorithms.
- Family Medicine and NCDC Weather Data: used by the UW Family Medicine department to provide physicians access to data that is collected and maintained independently by several clinics and also weather data from National Climate Data Center.
- Soil Sciences Classification: the BOREAS field experiment.

32

Visualization Model (1/2)

- It is based on mapping each source data record to a visual symbol on screen. "Plotting the data record" on some sort of graph.

[standard 2D/3D displays]
- Source data called TData (tabular data)
- GData (graphical data) is the visualization with attributes x, y, size, color, etc.
- Mapping: a function that produces a GData record from a TData record. This is data-independent. Only depend on the TData schema (table column headings, variable types of the columbs)
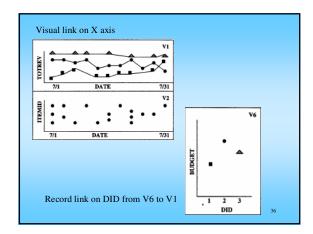
33

Visualization Model (2/2)

- View: the basic display unit in DEVise, consists of 3 layers: background, data display, and cursor display. Background and cursor display are data-independent.
- Each view has a mapping, TData, and a visual filter.
- A visual filter is a set of selections on the GData attributes. E.g. a range of x and y. A visual filter is ultimately translated to a query
- VGData: visible GData. This is computed from TData and is the data-dependent portion of a view.
- View template: the data-independent portion

34

Coordination views (1/2)

[Interactive linking and brushing]
- 2 mechanisms: Cursors and links
- A cursor allows the visual filter of one view (source view) to be seen as a high light in another view destination view). This is bi-directional.
- Visual link: visual filters of two views have share attributes. E.g. visual link on the x axis.
- Record link (positive or negative): a set of common TData attributes. The projection of the VGData on the linked attributes of the first linked view (the master) acts as a filter on the TData of the second linked view (the slave).

35

Visual link on X axis

Record link on DID from V6 to V1

36

Coordination views (2/2)

- Operator link: an operator (such as union, intersection) is applied to VGData(s) of link masters and creates a TData for the link slave.
- Aggregate link: the second view visualizes some aggregate function, e.g., sum and average.

37

Another Matrix reference!
"Operator Link" – Matrix Reloaded

LINK
THE MATRIX: RELOADED
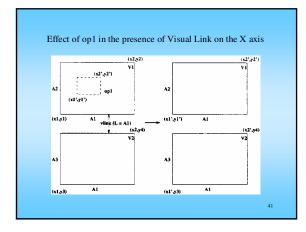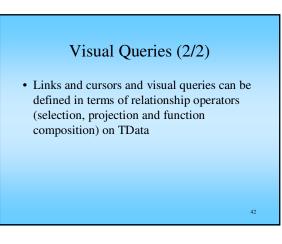FOX STUDIOS AUSTRALIA, OCTOBER 2001   WWW.THEMATRIX.COM

Organizing complex visual presentations

- A windows: collection of views together with the set of cursors and links
- A visual presentation: a collection of windows plus a collection of links and cursors.
- A visual template: the data-independent portion of a visual presentation.

39

Visual Queries (1/2)

1) op1: changing the x-y ranges.
2) op2: click and display the actual TData record
3) op3: Move a cursor
- A query (called a linked query) maybe be generated as a side-effect of a visual query.

40

Effect of op1 in the presence of Visual Link on the X axis

41

Visual Queries (2/2)

- Links and cursors and visual queries can be defined in terms of relationship operators (selection, projection and function composition) on TData

42

Example: Visual links on attribute L

$$vlink(v_1, v_2, L) \Rightarrow$$
$$\sigma^{G_1} = \sigma^{G_1}_{1-L} \circ \sigma^{G_1 2}_{L} \; and$$
$$\sigma^{G_2} = \sigma^{G_2}_{1-L} \circ \sigma^{G_1 2}_{L}$$

43

Visual Queries and SQL (1/3)

- Allows users who are not database experts to generate sophisticated SQL queries through intuitive graphical operations.
- Let T be a set of TData records (latitude, longitude, orders, totalamount)
- View 1 has a mapping that gives a scatter plot of totalamount vs. latitude.
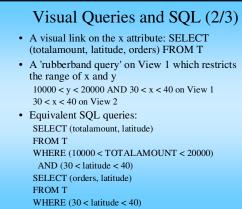- View 2 has a mapping that gives a scatter plot of order vs. latitude.
- The equivalent SQL queries are
- SELECT (totalamount, latitude) FROM T
- SELECT (order, latitude) FROM T

44

Visual Queries and SQL (2/3)

- A visual link on the x attribute: SELECT (totalamount, latitude, orders) FROM T
- A 'rubberband query' on View 1 which restricts the range of x and y

  10000 < y < 20000 AND 30 < x < 40 on View 1
  30 < x < 40 on View 2
- Equivalent SQL queries:

  SELECT (totalamount, latitude)
  FROM T
  WHERE (10000 < TOTALAMOUNT < 20000)
   AND (30 < latitude < 40)
  SELECT (orders, latitude)
  FROM T
  WHERE (30 < latitude < 40)

45

Visual Queries and SQL (3/3)

- Vice versa, an SQL query can be expressed using a visual presentation.
- Queries can operate on both local and remote data sources. This is exploited by DEVise.
- Evaluate query at remote sites if supported
- Otherwise retrieve complete relations and do the rest locally.

46

Advanced Exploration Tasks (1/2)

Integrated Access to Data and Metadata

- When datasets are very large and too much information is lost by compression, a powerful paradigm is to let users create summaries of data and to browse the summaries.
- E.g. statistical measures over subsets of the data. Support is built directly into the current version of DEVise.

47

Advanced Exploration Tasks (2/2)

Collaborative Analysis

- A user can save a visual template (the data-independent part) and send it to another user. Such a visual template is called an "active report".
- Future work: Share a visual representation and changes made by one user are automatically seen by all users.

48

Critiques

+ Well developed and evolving system with a lot of real applications and many feedback from domain experts

+ I like visual querying of large database that doesn't fit in main memory and then displaying the result visually.

- The simple x-y plot and bar graph are limiting.

- A visual presentation with 6 windows and 10 views in total might be disorienting.

49

50