Data Mining:

Concepts and Techniques

- Chapter 2 -

Ali Shakiba

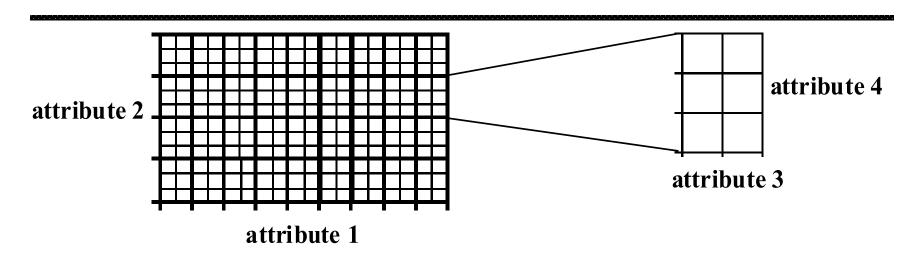
Vali-e-Asr University of Rafsanjan

based on slides by Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign Simon Fraser University ©2011 Han, Kamber, and Pei. All rights reserved.

Hierarchical Visualization Techniques

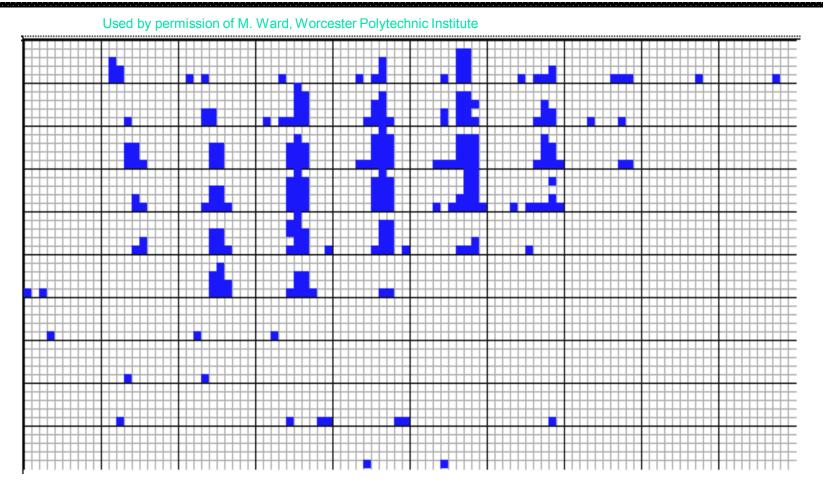
- Visualization of the data using a hierarchical partitioning into subspaces
- Methods
 - Dimensional Stacking
 - Worlds-within-Worlds
 - Tree-Map
 - Cone Trees
 - InfoCube

Dimensional Stacking



- Partitioning of the n-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately

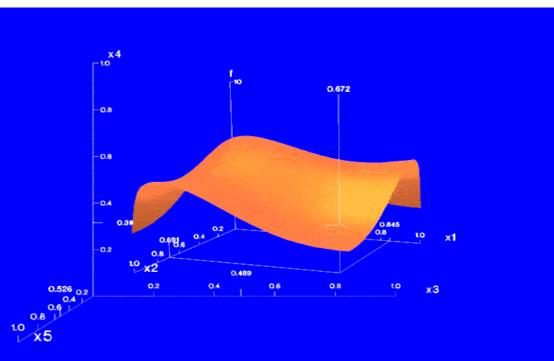
Dimensional Stacking



Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

Worlds-within-Worlds

- Assign the function and two most important parameters to innermost world
- Fix all other parameters at constant values draw other (1 or 2 or 3 dimensional worlds choosing these as the axes)
- Software that uses this paradigm
 - N-vision: Dynamic interaction through data glove and stereo displays, including rotation, scaling (inner) and translation (inner/outer)
 - Auto Visual: Static interaction by means of queries

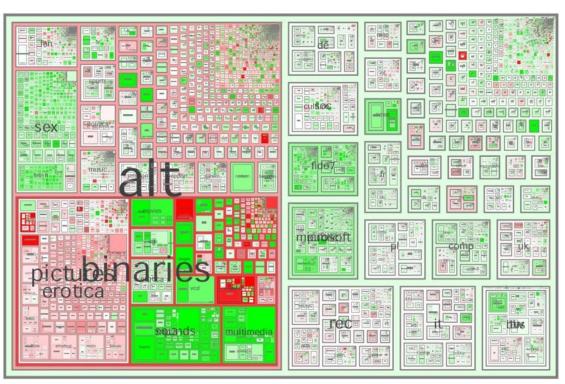


Tree-Map

- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values
- The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)

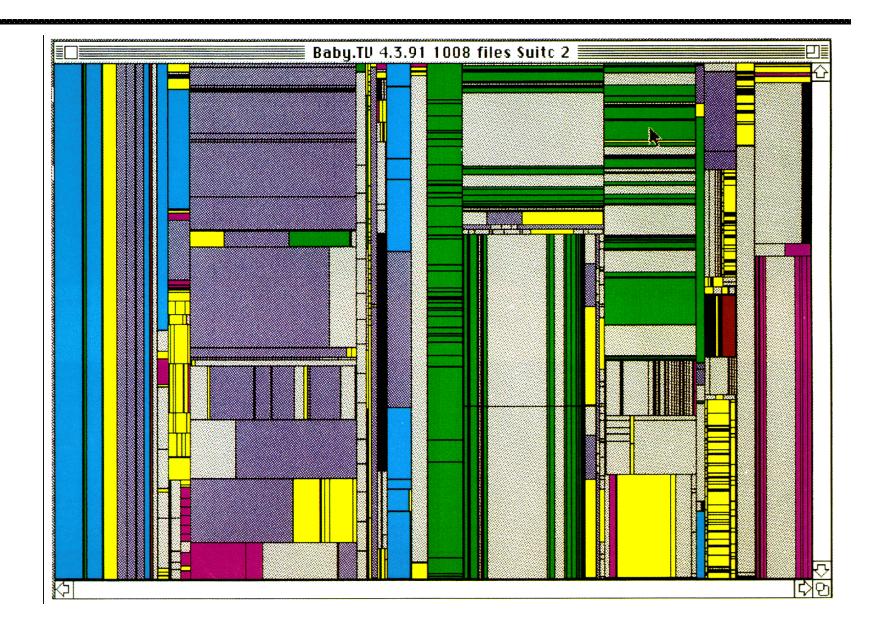
More at <u>https://en.wikipedia.org/wik</u> i/Treemapping

MSR Netscan Image



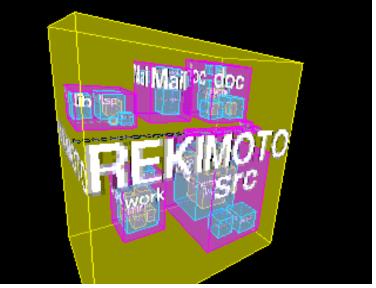
Ack.: http://www.cs.umd.edu/hcil/treemap-history/all102001.jpg 41

Tree-Map of a File System (Schneiderman)



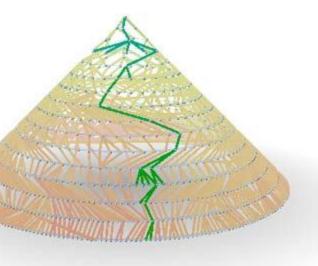
InfoCube

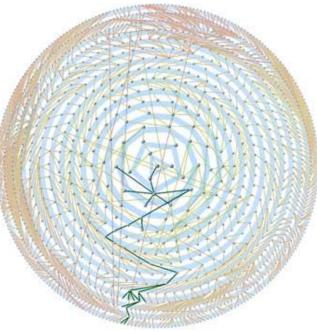
- A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes
- The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on



Three-D Cone Trees

- 3D cone tree visualization technique works well for up to a thousand nodes or so
- First build a 2D circle tree that arranges its nodes in concentric circles centered on the root node
- Cannot avoid overlaps when projected to 2D
- G. Robertson, J. Mackinlay, S. Card. "Cone Trees: Animated 3D Visualizations of Hierarchical Information", ACM SIGCHI'91
- Graph from Nadeau Software Consulting website: Visualize a social network data set that models the way an infection spreads from one person to the next





Ack.: http://nadeausoftware.com/articles/visualization

Visualizing Complex Data and Relations

- Visualizing non-numerical data: text and social networks
- Tag cloud: visualizing user-generated tags
 - The importance of tag is represented by font size/color
- Besides text data, there are also methods to visualize relationships, such as visualizing social networks

Penta on Denies	g 'M All Sm	Charges other Of lokescree	Harman Called Kind to Iraqis at Sentencing	Ruling wine shipme could u new ma	nts ncork	sky 'gui		Fev pito	/er	Missed cut not all bad for Tigei
Desec	r Ex	anted Cub ile Surfac US	es Tries to Hold on to Office	Stormy hurrican e season	Morgan St Grim over Costs		United Air inks labour deal with mechanics	Daily Playoff Essenti	Gi	erby Winner acomo Has enty to ove
ation	Harva Presic Piedg Million Wome Progra	lent Highway B es \$50 Exceeds B n for Limit (Upda en?s	ill, ush's flight from ate2 Italy diverted	predicte d	Toyota to Make Hybrids in US	Prospective Ar merger dominates annual mee	Alerts on Wireless Phones	The Golden Bear	to join and	and the second
Senate near battle on Bush nominees	rs Geory man not g to	gia missing pleads kids uilty b ^e says workers' tail	rimate asks Hob Seen of to donate Iver Ilver Slot Gay BS stat lady b marrian Valedictoria	Saudi leader: Kingdom can meet oil demand	Stocks rise after Treasury warns China	US warns China, says no forex violation 504m betting shop deal	Ander segnition Mr Transco Contrars & Anomen Stern Transco Construction Transco Transc	Cole Insists Gumers Have the Edge	Apologizes For Real Remark Byrd Takes Off In Nats Debut	a nd Canas A sepantina nac A sepantina A s
Troops, Militants Battle In	Ethiopia Opposit Governr Victory is Prem	ion: work nent Kidn Claim - for	apped in Oppose ul is well Oppose eign Gaza	Lucas (Indiana IV		Bi Si by	ocial Worker: oy Denied exual Abuse y Jackson	Nintenc o Plans a ?Revo		p compute emails
Mosul	UK Heipline Gets 320 Calls, 70 Emails About Lost	Kuwaiti women cabinet posts: speaker	Spenish Isymakers back plan tor talks with Baseue group ETA Lizbok Masternatio Plane	Kylie Minogue Battling Cancer	CBS 'Raymond finale draws 32 million viewers	r chap chap shoes	roud of Jinny Notics	Iution? in ?06 Experts Deb Study on Fat	Bittle of the Next Generation Format	
mediation ic	Piano Man 67 Dead And 35 Missing In Bangladesh Ferry Mishap	Its Grackdewn Was Far More Deadly Reclam Sinno	Libbek Masherrita (Markovani Karlovani Karlova	Movies still in slump	Firefigh er Tom is Sole Survivo	Projection Narriuge Property	ADC Subarning Responses Harry South States Harry St	New study suggests kudz helps curb bin drinking	er funde for health research	Patent Spits to brief state
Tuesday May 17, 200 ARCHVED UEB 180 PR 181 0000 0000		HOW I	LESS THAN TO MINUTES AGO HORE THAN TO MINUTES AGO MORE THAN TO MINUTES AGO	+ SELECT ALL CATE	GORIES V HATION	BUSINE	SS Z TECHHOLOG		LAYCUT: 🖃 SI	QLARIFIED STAAC

Newsmap: Google News Stories in 2005

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Similarity and Dissimilarity

Similarity

- Numerical measure of how alike two data objects are
- Value is higher when objects are more alike
- Often falls in the range [0,1]
- Dissimilarity (e.g., distance)
 - Numerical measure of how different two data objects are
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

- Data matrix
 - n data points with p dimensions
 - Two modes

 $\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$

- Dissimilarity matrix
 - n data points, but registers only the distance
 - A triangular matrix
 - Single mode

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Method 1: Simple matching
 - m: # of matches, p: total # of variables

$$d(i,j) = \frac{p-m}{p}$$

- Method 2: Use a large number of binary attributes
 - creating a new binary attribute for each of the M nominal states

Proximity Measure for Binary Attributes

 A 	contingency	table for	binary	data
-----------------------	-------------	-----------	--------	------

- Distance measure for symmetric binary variables:
- Distance measure for asymmetric binary variables:
- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

			oject <i>j</i>	
ta		1	0	sum
Object	, 1	q	r	q+r
Object	0	s	t	s+t
	sum	q+s	r+t	p
d	(i, j) =		$\frac{+s}{+s+t}$	
C	l(i, j) =	$=\frac{r}{q+q}$	$\frac{+s}{r+s}$	
sin	n _{Jaccaro}	l(i, j) :	$=\frac{q}{q+r}$	$\frac{l}{r+s}$

Dissimilarity between Binary Variables

• Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	Μ	Y	N	Р	N	N	N
Mary	F	Y	Ν	Р	N	Р	Ν
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

$$d (jack , mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d (jack , jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d (jim , mary) = \frac{1+2}{1+1+2} = 0.75$$

Standardizing Numeric Data

• Z-score:
$$Z = \frac{x-\mu}{\sigma}$$

- X: raw score to be standardized, μ: mean of the population, σ: standard deviation
- the distance between the raw score and the population mean in units of the standard deviation
- negative when the raw score is below the mean, "+" when above
- An alternative way: Calculate the mean absolute deviation

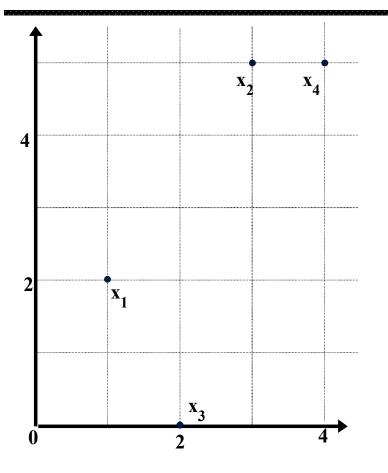
$$s_{f} = \frac{1}{n} (|x_{1f} - m_{f}| + |x_{2f} - m_{f}| + \dots + |x_{nf} - m_{f}|)$$

where

$$m_f = \frac{1}{n} (x_{1f} + x_{2f} + \dots + x_{nf})$$
 $x_{if} - n$

- standardized measure (*z*-score): $Z_{if} = \frac{T_{if}}{S_{c}}$
- Using mean absolute deviation is more robust than using standard deviation

Example: Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
<i>x1</i>	1	2
<i>x2</i>	3	5
<i>x3</i>	2	0
<i>x4</i>	4	5

Dissimilarity Matrix

(with Euclidean Distance)

	<i>x1</i>	<i>x2</i>	<i>x3</i>	<i>x4</i>
<i>x1</i>	0			
<i>x2</i>	3.61	0		
<i>x3</i>	5.1	5.1	0	
<i>x4</i>	4.24	1	5.39	0

Distance on Numeric Data: Minkowski Distance

Minkowski distance: A popular distance measure

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h} + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h$$

where $i = (x_{i1}, x_{i2}, ..., x_{ip})$ and $j = (x_{j1}, x_{j2}, ..., x_{jp})$ are two *p*-dimensional data objects, and *h* is the order (the distance so defined is also called L-*h* norm)

- Properties
 - d(i, j) > 0 if $i \neq j$, and d(i, i) = 0 (Positive definiteness)
 - d(i, j) = d(j, i) (Symmetry)
 - $d(i, j) \le d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a metric

Special Cases of Minkowski Distance

- h = 1: Manhattan (city block, L₁ norm) distance
 - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

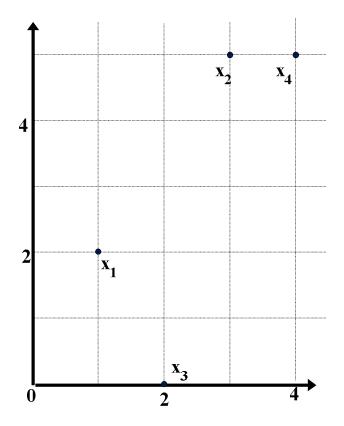
- h = 2: (L₂ norm) Euclidean distance $d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$
- $h \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
 - This is the maximum difference between any component (attribute) of the vectors

$$d(i, j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Example: Minkowski Distance

Dissimilarity Matrices

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L₁)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L₂)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_{∞}	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0