### 9. Text & Documents

Visualizing and Searching Documents

Dr. Thorsten Büring, 20. Dezember 2007, Vorlesung Wintersemester 2007/08





Maximilians-Universität\_\_\_\_ München\_\_\_\_

### Outline

- $\equiv$  Characteristics of text data
- $\equiv$  Detecting patterns
  - ≣ SeeSoft
  - ∃ Arc diagrams
  - ∃ Visualizing Plagiarism

#### ■ Keyword search

- ∃ TextArc
- $\equiv$  Enhanced scrollbar
- TileBars
- Cluster Maps
  - $\equiv$  Visualization for the document space
  - WEBSOM
  - ∃ ThemeScapes
- $\equiv$  Cluster map vs keyword search



München

**Text & Documents** 

- $\equiv$  The main mean to store information
- $\equiv$  Huge existing resources: libraries, WWW
- $\equiv$  What to visualize?
- $\equiv$  Text is of nominal data type, but with many additional and interesting properties
- Text structure
- 📃 Meta data
  - Author
  - Dates
  - Descriptions
- $\equiv$  Relations between documents (e.g. citation, similarity)
- $\equiv$  Relevance of documents to a query
- Text statistics (e.g. frequency of different words)
- $\equiv$  Content / Semantics



Universität\_\_\_ München\_\_\_\_

Outline

- $\equiv$  Characteristics of text data
- Detecting patterns
  - ≣ SeeSoft
  - Arc diagrams
  - ∃ Visualizing Plagiarism
- Keyword search
  - ∃ TextArc
  - $\equiv$  Enhanced scrollbar
  - TileBars
- Cluster Maps
  - $\equiv$  Visualization for the document space
  - WEBSOM
  - ∃ ThemeScapes
- $\equiv$  Cluster map vs keyword search



Universität\_\_\_\_ München\_\_\_\_\_

SeeSoft

- Eick et al. 1993
- Software visualization tool to display code line statistics (e.g. age, programmer, number of execution in recent test, etc.)
- Encoding
  - $\equiv$  Each column represents a file
  - $\equiv$  Height of column: length of document
  - Files exceeding the height of the screen are continued over to the next columns
  - $\equiv$  Each row represents a line of code
  - $\equiv$  Width of row: length of line
  - $\equiv$  Color: age of the line (red: newest; blue: oldest)

www.medien.ifi.lmu.de

- $\equiv$  Scales up to 50,000 lines on a single screen
- $\equiv$  Example: 20 files with 9,365 lines of code
- Reading windows controlled by virtual magnifying boxes

LMU Department of Media Informatics

1 121211

thorsten.buering@ifi.lmu.de



München

# Arc Diagrams

- Wattenberg 2002
- $\equiv$  Visualizes repetition in string data
- $\equiv$  Application domains: text, DNA sequences, music
- Approach: to avoid clutter, only visualize an essential subset of all possible pairs of matching substrings
- $\equiv$  Display string on a single line
- $\equiv$  Connect the consecutive intervals by a semi-circular arc
  - $\equiv$  Thickness of the arc: length of the matching substring
  - $\equiv$  Height of the arc: proportional to the distance of substrings

### 28746391479735648274639137



München

# Arc Diagrams

- $\equiv$  Apply translucency to not obscure matches
- Still: for strings with a high frequency of small repeated substrings the visualization may cause clutter
- $\equiv$  Provide users with the ability to filter by minimum substring length to consider







Maximilians– Universität\_\_\_\_ München\_\_\_\_\_

### Arc Diagrams

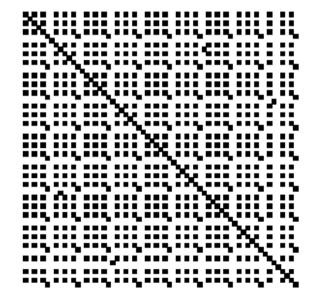
- $\equiv$  Comparison to a dotplot diagram
- Recap Matrix diagram
  - $\equiv$  Correlation matrix
  - $\equiv$  String of n symbols  $a_1$ ,  $a_2$ ,  $a_n$  is represented by an n\*n matrix
  - $\equiv$  Pixel at coordinate (i, j) is black if  $a_i = a_j$
  - $\equiv$  Can handle very large datasets
  - $\equiv$  Shows both small and large-scale structures
- Heavy clutter caused by small substrings with high frequency: n repetitions of a substring lead to  $n^2$  visual marks
- $\equiv$  Arc Diagrams mark only similar substrings, which are subsequent



LMU Department of Media Informatics

thorsten.buering@ifi.lmu.de

Slide 8 / 37

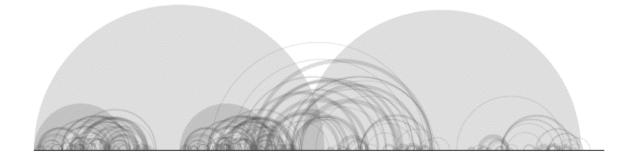




Universität\_\_\_\_ München

# Arc Diagrams

- $\equiv$  Applied to music, Minuet in G Major, Bach
- $\equiv$  Shows classic pattern of a minuet: two main parts, each consisting of a long passage played twice
- $\equiv$  Parts are loosely related: bundle of thin arcs connecting the two main parts
- $\equiv$  Overlap of the two main arcs shows that the end of the first passage is the same as the beginning of the second passage





München

- Ribler & Abrams 2000
- $\equiv$  Problem: programming assignment in a class with large number of students
- $\equiv$  High probability of plagiarism
- $\equiv$  Need to compare every document (code file) with every other document
- $\equiv$  Visualization must support two steps
  - $\equiv$  Highlight suspicious documents
  - Allow for detailed examination of the similar passages high level of similarity between documents may not be due to cheating (e.g. headers)

- Categorical Patterngram
- Visualize frequencies of sequences of characters present in more than one document
- Remove all non-printable characters in the document collection
- Define length of character sequence to analyse (in the example: 4)
- ∃ Histogram-like approach
  - $\equiv$  X-axis: start character of sequence
  - $\equiv$  Y-axis: number of documents containing the sequence
  - $\equiv$  Doc at Y = 1: base document to compare against all other documents



Toy0: This is a test.

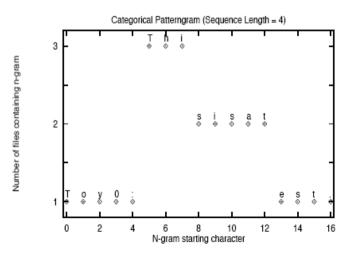
Figure 1. Toy File 0

Toy1: Oh yes. This is a test too.

Figure 2. Toy File 1

Toy2: Toy2 has little in common with the other two. This is common.

### Figure 3. Toy File 2



- E Composite Categorical Patterngram
- $\equiv$  Visualizes which particular documents are similar
- Y-axis: each value corresponds to an individual document



Toy0: This is a test.

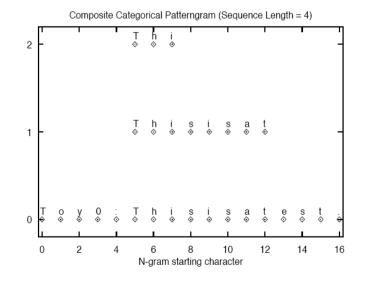
Figure 1. Toy File 0

Toy1: Oh yes. This is a test too.

Figure 2. Toy File 1

Toy2: Toy2 has little in common with the other two. This is common.

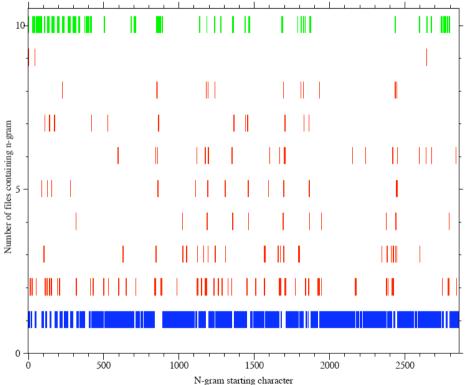
### Figure 3. Toy File 2





- Case study
- Students were asked to extend a sample program of about 30 lines of code

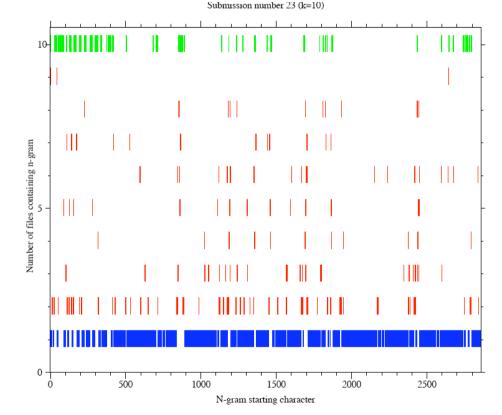
- Average completed program was about 150 lines 150 lines Submission via email Graphic shows categorical patterngram for a single submission  $\equiv$  Sequence length = 10
  - Sequence length = 10
  - Lines not text due to high density
  - Rather confusing color coding
- Color coding (not very reasonable)
  - Green: frequency  $\geq 10$
  - Red: frequency < 10
  - Blue: base document
- Plagiarism or not?



Submission number 23 (k=10)



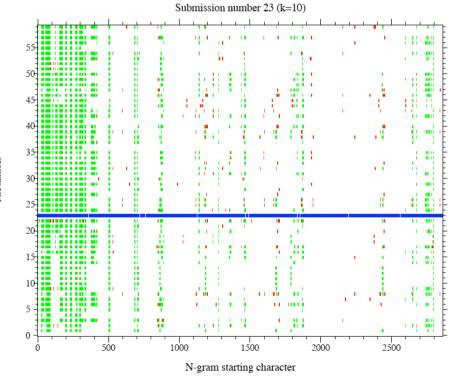
- $\equiv$  What to look out for?
  - E Sequences that occur frequently are not of interest - all points with  $y \ge 10$  are plotted as y = 10
  - Suspicious: accumulation of points with low frequencies
- Analysis
- $\equiv$  Majority of points are plotted at Y = 1
- ➡ Hence most 10-char sequences are unique to the base document
- Number of points plotted at Y = 2, but evenly distributed





Maximilians– Universität\_\_\_ München

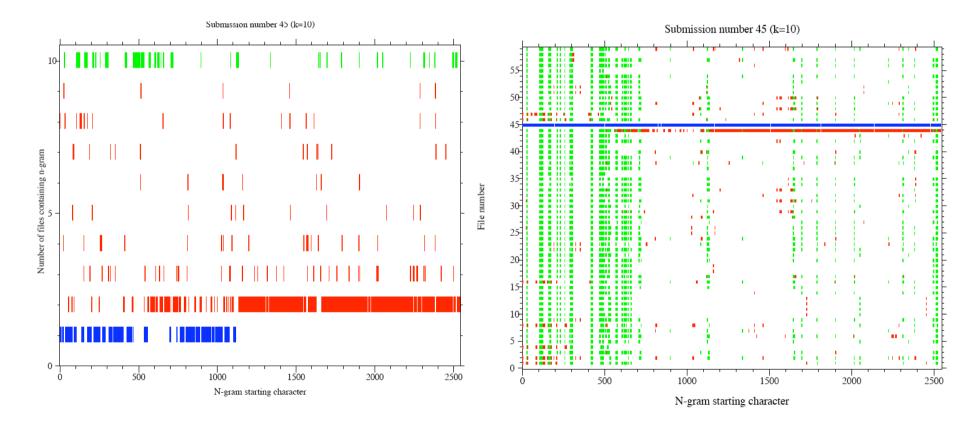
- E Composite Categorical Patterngram for the submission
- Solid line represents the base document (submission number 23)
- Example a Large number of points plotted in the range of  $x_{\frac{1}{2}}$ = [0; 500]: email message header
- Other frequent sequences due to the sample program
- $\equiv$  Pattern typical for independent work





Ludwig—— Maximilians– Universität— München——



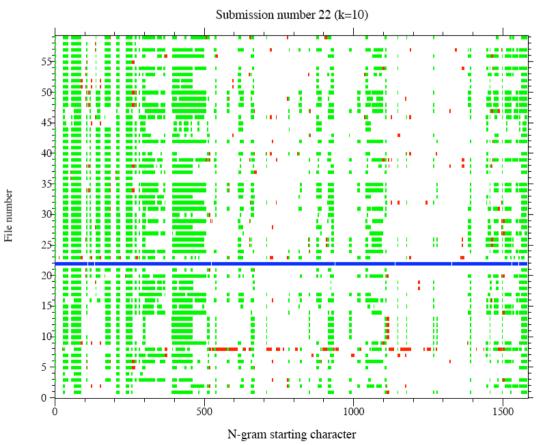




Maximilians– Universität\_\_\_\_ München\_\_\_\_

# Visualizing Plagiarism

 $\equiv$  Patterngram of more subtle plagiarism



www.medien.ifi.lmu.de

thorsten.buering@ifi.lmu.de



Slide 18/37

München

- $\equiv$  What may a student do to mask plagiarized code
- $\equiv$  Change variable names
- Minimize masking effect by replacing all alphanumeric strings in all documents into single characters
- Two documents with the same code but different variable names will produce identical patterngrams



Maximilians-Universität\_\_\_\_ München\_\_\_\_

### Outline

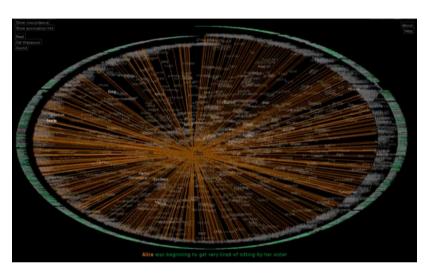
- $\equiv$  Characteristics of text data
- $\equiv$  Detecting patterns
  - ≣ SeeSoft
  - Arc diagrams
  - ∃ Visualizing Plagiarism
- Keyword search
  - ∃ TextArc
  - $\equiv$  Enhanced scrollbar
  - TileBars
- Cluster Maps
  - $\equiv$  Visualization for the document space
  - WEBSOM
  - ∃ ThemeScapes
- $\equiv$  Cluster map vs keyword search

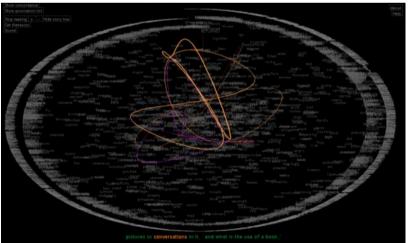


München

### **TextArc**

- http://www.textarc.org/ demo
- Represents the entire text as 1 pixel lines in an outer circle
- $\equiv$  Text is revealed via mouse-over
- $\equiv$  Words are repeated in inner circle at a readable size
- Position of the words depend on where the word appears in the document
- Words that appear throughout the novel will be drawn to the center
- $\equiv$  Frequent words stand out
- $\equiv$  Example visualizes the novel "Alice in Wonderland"
- $\equiv$  Various visualization features
  - $\equiv$  Association of words
  - $\equiv$  Word frequency
  - $\equiv$  Reading order of words (animated)







Maximilians– Universität\_\_\_\_ München

# Search Terms on a Scrollbar

### ■ Byrd 1999

- Searching of keywords in a single document
- Color coding to map each occurrence of a keyword in the document as a small colored icon in the scrollbar
- Provides an overview of the entire document, not only of the portion currently visible
- Users can directly jump to keyword occurences by moving the slider thumb

S TITLE NOT FOUND	_ 🗆 ×
Relevant     Not Relevant     Document 1D: 19912	
nealth and sarety or starr and inmates, the bureau or Prisons will restrict areas and circumstances in which smoking is permitted within its institutions and offices.	1
(a) All areas of Bureau of Prisons facilities and vehicles are no smoking areas unless specifically designated as smoking areas by the Chief Executive Officer consistent with the guidelines set forth in this rule.	1
(b) Chief Executive Officers shall limit smoking areas to a minimum number of locations, consistent with effective operations. Under no circumstances shall smoking be permitted in the following areas, except as noted in &Section 551, 162(a):	2
<ol> <li>Elevators,</li> <li>Storage Rooms and Warehouses.</li> <li>Libraries.</li> </ol>	
<ul> <li>(4) Corridors and Halls,</li> <li>(5) Dining Facilities,</li> <li>(6) Kitchen and Food Preparation Areas,</li> </ul>	1
<ul> <li>(7) Medical/Dental Care Delivery Areas,</li> <li>(8) Institution/Government Vehicles.</li> <li>(9) Administrative Areas and Offices,</li> <li>(10) Auditoriums.</li> </ul>	7
<ul> <li>(10) Auditoriums,</li> <li>(11) Class and Conference Rooms,</li> <li>(12) Gymnasiums and Exercise Rooms, and</li> <li>(13) Restrooms.</li> <li>SSection, 551.161</li> <li>Definition.</li> </ul>	-
For purpose of this rule, smoking is defined as carrying or inhaling a lighted cigar, cigarette, pipe or other lighted tobacco products.	
anti 📟 smoke 🔜 action 🤤 govern	



Maximilians-Universität\_\_\_ München

### **TileBars**

### 

- Problem with document ranking of common search engines?
- $\equiv$  Ranking approach is opaque:
  - What role did the query terms played in the ranking process
  - What is the relationship between the query terms in the document
- TileBars attempts to le the users make informed decisions about which documents and passages to view

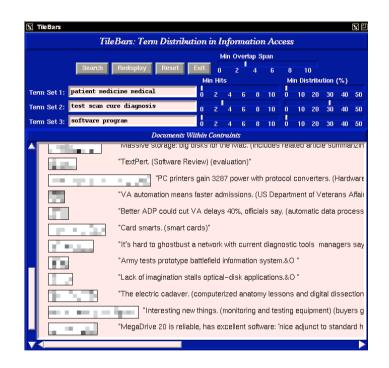




Ludwig Maximilians-Universität München

### **TileBars**

- $\equiv$  Users provide sets of query terms
  - $\equiv$  OR within a set
  - AND between sets
- Documents are partitioned into adjacent, non-overlapping multi-paragraph segments
- Each document of the result set is represented by a rectangle width indicates relative length of the document
- $\equiv$  Stacked squares correspond to text segments
- $\equiv$  Each row of the stack corresponds to a set of query terms
- Darkness of the square indicates the frequency of terms from the corresponding term set - (Why is this a reasonable color mapping?)
- $\equiv$  Title + initial words appear next to each document
- Users can click on segments to retrieve the corresponding text





Ludwig—— ' Maximilians– Universität— München——

### **TileBars**

### ■ Analysis hints

- Overall darkness indicates that all term sets are discussed in detail throughout the document
- When terms are discussed simultaneously the tiles blend together causing an easy to spot block
- Scattered term set occurrence show large areas of white space
- Helps to distinguish between passing remarks and prominent topic terms
- Users may also set distribution constraints to refine the query
  - $\equiv$  Minimum number of hits per term set
  - Minimum distribution (percentage of tiles containing at least one hit)
  - $\equiv$  Minimum adjacent overlap span

ХТ	ile Bars															又四
		Ti	ileBars: Term	Distrib	ution	in I	nfo	rma	tior	ı Ac	cess	;				
							Min (	)verl:	ap Sj	pan						
		Search	Redisplay	Reset	Exit	0		2	4	6	8		0			
Tow	n Set 1:	natient me	dicine medical			in Hi								ition		
	n Set 2:		cure diagnosis			2	4	6	8	10	0		20	30 I	40	50
		software p				2	4	6	8	10	0	10	20	30	40	50
Tem	n Set 3:	soltware p	<u> </u>	ocuments	0	2	4	6	8	10	0	10	20	30	40	50
	Total: 9		De	ocumenis	wunin	com	irain	15								
	rotai. 5															
	100		"Paragon po	irts its 2–	D imag	ing :	softv	/are 1	to Di	EC's	AVS	S. (Pa	arago	on Irr	nagir	ig In
			"DOD health	i system i	costs k	real	<\$1.	1 bill	ion t	parrie	er. (C	omp	osite	e Hea	lth C	are
		"Computer graphics in medicine: from pictures to analysis.&O (Industry Profi														
		"Lab∀IEW 2.0 still just over the horizon. (National Instruments Lab∀IEW 2.0														
		"Hospital bridges islands of data. (includes related articles on growth of a me														
		"VA automation means faster admissions. (US Department of Veterans Affair														
		"Better ADP could cut VA delays 40%, officials say. (automatic data process														
	100	"Army tests prototype battlefield information system.&O "														
		"The electric cadaver. (computerized anatomy lessons and digital dissection														
	(															



Maximilians-Universität\_\_\_\_ München\_\_\_\_

### Outline

- $\equiv$  Characteristics of text data
- $\equiv$  Detecting patterns
  - ≣ SeeSoft
  - ∃ Arc diagrams
  - ∃ Visualizing Plagiarism
- Keyword search
  - ∃ TextArc
  - $\equiv$  Enhanced scrollbar
  - TileBars
- Cluster Maps
  - $\equiv$  Visualization for the document space
  - WEBSOM
  - ∃ ThemeScapes
- $\equiv$  Cluster map vs keyword search



München

### **Cluster Maps**

- $\equiv$  Downscaling of n-dimensional document space to 2D
- $\equiv$  Map of a document collection
- $\equiv$  Similar documents are placed close to each other
- $\equiv$  Dissimilar documents are placed farer apart from each other
- Provide thematic overview for exploration (same concept as product arrangements in a store)
- $\equiv$  How to Vector space model and map construction
  - $\equiv$  Create inverted index of document collection
  - Exclude stop words and the most frequent words ("and" may not be a good discriminator of content)
  - $\equiv$  Matrix of indexing words versus documents gives you document vectors
  - $\equiv$  A document vector reflects the frequency of index words occurring in the document



### **Cluster Maps**

- ➡ How to Vector space model and map construction (continued)
  - Compute similarity between pairs of documents (e.g. dot product of vectors)
  - $\equiv$  Layout documents in 1D/2D/3D
- $\equiv$  Common approaches
  - $\equiv$  Spring model of graph layout
  - $\equiv$  Multi-dimensional scaling
  - $\equiv$  Clustering (e.g. hierarchical)
  - Self-organizing maps (SOM aka Kohonen map)

#### Document vectors

	Doc 1	Doc 2	Doc 3
"Artificial"	1	2	0
"Creativity"	2	1	0
"Java"	0	0	3

#### Similarity Matrix

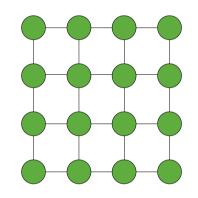
	Doc 1	Doc 2	Doc 3
Doc 1	1	0.66	0
Doc 2	0.66	1	0
Doc 3	0	0	1



Universität\_\_\_\_ München

### SOM

- $\equiv$  Unsupervised learning algorithm
- $\equiv$  SOM map is formed from a regular grid of neurons (nodes)
- $\equiv$  Each node has
  - $\equiv$  An x y coordinate in the grid
  - $\equiv$  A weight vector of the same dimensionality as the input vectors
- Input vectors
  - $\equiv$  Used to *train* the map
  - $\equiv$  Represent collection of objects
- $\equiv$  In case of visualizing text, input vectors are usually equal
  - to document vectors



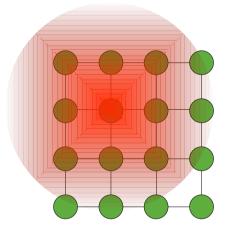
Network of 4x4 nodes

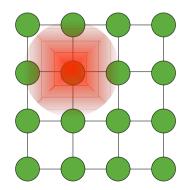
### Ludwig— Maximilians-

Universität\_\_\_ München

# SOM - Algorithm

- $\equiv$  1. Start with assigning small random weights to the nodes of the grid
- $\equiv$  2. Chose a vector at random from the set of input vectors and present it to the grid
- 3. For each node: calculate the Euclidean distance between each node's weight vector and the current input vector - the closest node is called the Best Matching Unit (BMU)
- 4. Calculate the radius of the BMU (radius diminishes with each time-step)
- 5. For each node within the radius of the BMU: adjust the weights to make them more similar to the input vector the closer a node is to the BMU, the more its weights get altered
- $\equiv$  6. Repeat step 2 for N iterations
- $\equiv$  When training is completed each document is assigned to its BMU







### **Cluster Maps**

### 📃 Lin 1992

- $\equiv$  Personal collection of 660 research documents
- $\equiv$  2500 learning iterations
- $\equiv$  Labeled word show most frequent title words
- $\equiv$  Size maps to frequencies of occurrence of the words
- Neighboring relationships of areas indicate frequencies of the co-occurrence of words

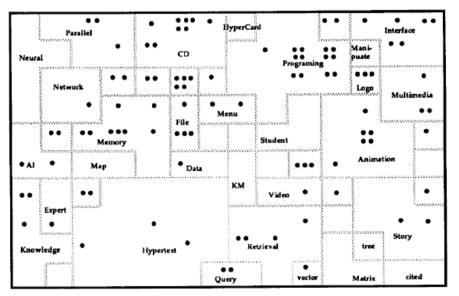
		Para	allel	c		н	perCa	r		objec		terfa 1	ice
Neur	al			, v	5						Mani- puate		
	Netv	vork	ttend	meta phor	cro: tive	Apple		Prog	nam		Logo		
						Men	ų						imedia
	theory	eory Memory			File Augu- ment		belle	Student		Animation		Bost tn	
AI		Mi	ap		Dat	a		Mana- gem.	Innov			Medi	
		elec- tron					КМ	Vi	deo	WA			Conf
Epert								Algo- eithm		Stor	y		
now	ledge		Hyj	pertex	t		Re	triev	al		tree		
	Medic					Que	ry		vecto	Ma	trix	ci	ted



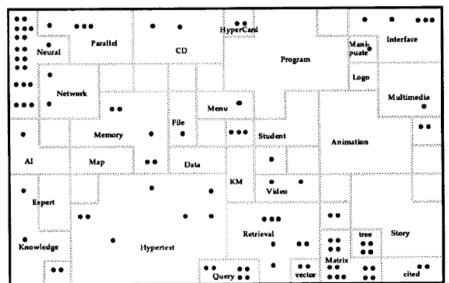
Maximilians– Universität<u></u> München

### **Cluster Maps**

 $\blacksquare$  Research interest changing over time



(a) Distribution of the first 100 documents in the personal collection







Maximilians-Universität\_\_\_ München

### WEBSOM

- http://websom.hut.fi/websom/
- SOM of Finnish news bulletins for exploring and retrieving documents
- $\equiv$  Labels show the topics of areas in the SOM
- Coloring encodes density light areas contain more documents
- $\equiv$  Navigation via zooming and panning
- Documents can be retrieved on the lowest level of the visualization
- 📃 Demo

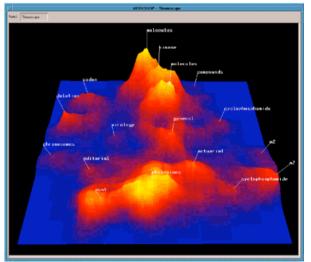
va donner	kuutusala <sup>lipp</sup>	onen tulva
moottoritie	lundh esiopetusuudi	olvi stus tulos
rahjan		örssi kannattavuus
	pomm. saku	L
<mark>rauhanpalkinto</mark>		ikinen
olvi linnainmaa alho	kausi	lakko
keittiö	kärkölässä	
loviis		isabet muotokuva
mainonta	schengen	lamminen
työttömyys cr karhu	eutzfeldt skp	albright
korea	ahallaan nii	jeltsin nistö



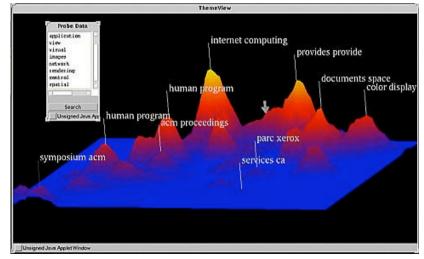
München

### **ThemeScapes**

- Wise et al. 1995
- $\equiv$  Map document density to third dimension
- $\equiv$  News article visualized as an abstract 3D landscape
- Mountains represent frequent themes in the document corpus (height proportional to number of documents relating to the theme)
- $\equiv$  Spatial characteristics of the map should map to interconnections of themes



http://nd.loopback.org/hyperd/zb/spire/spire.html LMU Department of Media Informatics www.



#### http://infoviz.pnl.gov/technologies.html

www.medien.ifi.lmu.de



Maximilians-Universität\_\_\_\_ München\_\_\_\_

### Outline

- $\equiv$  Characteristics of text data
- $\equiv$  Detecting patterns
  - ≣ SeeSoft
  - ∃ Arc diagrams
  - ∃ Visualizing Plagiarism
- Keyword search
  - ∃ TextArc
  - Enhanced scrollbar
  - TileBars
- Cluster Maps
  - $\equiv$  Visualization for the document space
  - WEBSOM
  - ∃ ThemeScapes
- $\equiv$  Cluster map vs keyword search



# Cluster Map vs Keyword Search

- E Chris North
- Cluster Map pros
  - **E** Facilitates non-targeted exploration and browsing by spatially organizing documents
  - Provides overview of document set: major themes, sizes of clusters, relationships between themes
  - $\equiv$  Scales up
- $\equiv$  Cluster Map cons
  - $\equiv$  How to label groups?
  - $\equiv$  What does the space mean? How to label space?
  - $\equiv$  Where to locate documents with multiple themes: both mountains, between mountains, ...?
  - $\equiv$  Relationships within documents?
  - $\equiv$  Algorithm (SOM) is time-consuming



# Cluster Map vs Keyword Search

- Chris North
- $\equiv$  Keyword search pros
  - $\equiv$  Reduces the browsing space according to user's interests
- $\equiv$  Keyword search cons
  - $\equiv$  What keywords do I use?
  - $\equiv$  What about other related documents that don't use these keywords?
  - $\equiv$  No initial overview
  - Mega-hit, zero-hit problem



München\_\_\_\_

**Additional Sources** 

- $\equiv$  Jonn Stasko, lecture material, CS 7450
- $\equiv$  Chris North, lecture material, CS 5764