Information Visualisation
What is Information Visualisation?

• Visualising *discrete* data with no or minor spatial information

• Visualisation of important information contained in abstract data types (trees, graphs)
  – Needs to be intuitive
  – Such that people can easily and quickly understand

• **Tools for**
  – *Extraction of information from the data*
  – *Discovery of new knowledge*
Overview

- **Information Visualisation**
  - Univariate, bivariate, trivariate, multi-variate data
  - Relations visualized by trees, and graphs
  - Document visualization
Data types

- Quantitative data
  - Univariate
  - Bivariate
  - Trivariate
  - Multivariate

- Relationship
  - Graphs
  - Trees
  - Mapping, figures

- Document Visualization
Univariate Data, Bivariate Data

• Data samples with one or two attributes
• Can use scatter plots, histograms
Trivariate Data

Scatterplots

Scatterplot matrix

Not clear if D is more expensive than B and C
Trivariate Data

Scatterplot Matrix: Visualizing the relations of every two variables
Multivariate Data

Parallel Coordinates

Star plots
Parallel Coordinates
Parallel Coordinates

Car data :
http://eagereyes.org/techniques/parallel-coordinates
Parallel Coordinates

Direct correlation
Parallel Coordinates

Inverse Relations
Brushing

Select some data using one of the coordinates

Brushing years 1980 to 1982
Brushing

Brushing the years 1970 to 1972
Limitations

Visual clutter

- Many lines cluttered together making it impossible to see anything
- Too many dimensions make things difficult to see
Solutions

• Re-ordering the axes to make the correlated coordinates adjacent to each other

• Negating the coordinates such that the inverse correlation becomes direct correlations

• Clustering the axes

We can use the biclustering algorithm that considers both the samples and the dimensions

Watanabe et al. 2015
K-Means Clustering

Computing clusters where the following criteria is minimized

$$
\sum_{i=1}^{n} \|r_i - \theta_{\kappa(i)}\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - \theta_{\kappa(i)j})^2.
$$

where

\{r_1, \cdots, r_n\}, \ r_i = (x_{i1}, \cdots, x_{id}) \in \mathbb{R}^d : \text{samples}

\theta_k = (\theta_{k1}, \cdots, \theta_{kd}) : \text{cluster means}

\{\kappa(i) \in \{1, 2, \cdots, K\}\}_{i=1}^{n} : \text{cluster labels}
Biclustering

Clustering the samples and the dimensions so that the correlated dimensions are joined into the same dimension cluster

\[
\sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - \mu_{\kappa(i),\lambda(j)})^2,
\]

\(\kappa(i) \in \{1, \cdots, K\} \ (i = 1, \cdots, n)\) : sample cluster

\(\lambda(j) \in \{1, \cdots, L\} \ (j = 1, \cdots, d)\) : dimension cluster
In this section, we describe the biclustering method, which clusters dimensions and data samples in the dataset, so that we can provide an efficient guide to the data analysts by allowing them to systematically explore the data. To achieve this, we develop a block modeling framework which maintains the spherical constraints during the optimization to deal with the correlation coefficient as the objective function (Section 4.2). Then, we describe a fundamental biclustering algorithm based on the block relation as well as positive correlation (Section 4.2). The biclustering algorithm is accomplished by incorporating the biclustering algorithm (Section 5), while the data exploration is composed of several views.

We provide two main functions, including outlier data samples for further exploration iteratively until reaching a satisfactory result. The present feature subspace extraction can intentionally generate a highly-correlated subspace and remove poorly-correlated data samples and green color represents highly-correlated function values, analysts can evaluate the goodness of the current subspace and data samples are simultaneously grouped through the process of automatically generating a highly-correlated subspace and removing poorly-correlated data samples based on the HSV model. Thanks to this color assignment, analysts can judge which cluster the current outlier data samples belong to easily.

Figure 2 shows the overall framework of the present approach. Figure 3 (a) illustrates the data matrix, and (b) presents a schematic representation of the block model for four dimensions, and cluster labels. Let us consider clustering of dimensions and data samples in the dataset, so that we can provide an automatic clustering step, where the highly-correlated dimensions and data samples are simultaneously grouped through the process of automatically generating a highly-correlated subspace and removing poorly-correlated data samples.

\[
\begin{align*}
M & = \{x_{ij} \} \\
& = \begin{pmatrix}
\begin{array}{cccc}
\hdots & \hdots & \hdots & \hdots \\
\vdots & x_{ij} & \hdots & \hdots \\
\vdots & \hdots & \hdots & \hdots \\
\end{array}
\end{pmatrix}
\end{align*}
\]

which shows that the original spherical k-means deals only with the positive correlation coefficient, which is suitable for PCP, and is in fact applied for sophisticated data analysis. Our approach begins with the spherical k-means clustering uses the following objective function, and conducts k-means clustering on a high-dimensional unit sphere \([4, 3]\). After describing the objective function of the spherical k-means that takes into account negative correlation, we will also introduce the spherical k-means with negative correlation coefficient, which is suitable for PCP, and is in fact applied for sophisticated data analysis. Our approach begins with the spherical k-means clustering uses the following objective function, and conducts k-means clustering on a high-dimensional unit sphere \([4, 3]\). After describing the objective function of the spherical k-means that takes into account negative correlation, we will also introduce the spherical k-means with negative correlation coefficient, which is suitable for PCP, and is in fact applied for sophisticated data analysis.
More results

(a) $K = 9$ and $L = 6$
Data types

- Quantitative data
  - Univariate
  - Bivariate
  - Trivariate
  - Multivariate

- Relationship
  - Graphs
  - Trees
  - Mapping, figures

Document Visualization
Visualizing Relations

Relation: A logical or natural association between two or more things; relevance of one to another; connection

Usually use lines to represent the relations
Tree visualization

Trees have hierarchical structures

No close loops

So many methods: see

http://vcg.informatik.uni-rostock.de/~hs162/treeposter/poster.html
Treemaps

Display hierarchical (tree-structured) data as a set of nested rectangles

The area representing a scalar attribute

The leaf nodes are often colored to visualize another attribute data
Worldmapper

http://sasi.group.shef.ac.uk/worldmapper/

Distorted maps according to numbers: Cartograms
Graph Visualization

Visualizing correlation of different nodes

E.g.

- social networks
- citation networks

vizster
Facebook relations
Facebook relations

“I defined weights for each pair of cities as a function of the Euclidean distance between them and the number of friends between them. Then I plotted lines between the pairs by weight, so that pairs of cities with the most friendships between them were drawn on top of the others. I used a color ramp from black to blue to white, with each line's color depending on its weight. I also transformed some of the lines to wrap around the image, rather than spanning more than halfway around the world.”
Formal Aesthetics

Minimize node-node / node-edge occlusions
Minimize edge crossings
Minimize edge bends
Maximize symmetry
Maximize the minimum angle between edges
Maximize edge orthogonality
Formal Aesthetics Metrics

Minimize edge crossings

Minimize edge bends
Formal Aesthetics Metrics

Maximizing symmetry

Maximizing the minimum angle between edges leaving a node
Another Application

metro maps : Wu et al. 2013

Figure 1: Wheelchair accessible stations in Prague metro. (a) Octilinear and (b) orthogonal annotated maps.
Data types

- Quantitative data
  - Univariate
  - Bivariate
  - Trivariate
  - Multivariate

- Relationship
  - Graphs
  - Trees
  - Mapping, figures

- Document Visualization
Document Visualisation

- **Motivation:**

<table>
<thead>
<tr>
<th>Action</th>
<th>Units of Information transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typing at 10 bytes per second</td>
<td>1</td>
</tr>
<tr>
<td>Mouse Operations.</td>
<td>2</td>
</tr>
<tr>
<td>Reading</td>
<td>3-40</td>
</tr>
<tr>
<td>Hearing</td>
<td>60</td>
</tr>
<tr>
<td>Visualisation and Pattern Recognition</td>
<td>12,500</td>
</tr>
</tbody>
</table>

- Visualisation is considerably faster than hearing / reading!

Source: Silicon Graphics Inc.
Visualisation of Documents

• Motivation: large bandwidth of human visual system
  - 100s millions of documents available on-line
  - information only in textual form

• ‘Visualising the non-visual’
  - searching for scientific papers
  - analysing witness statements
  - awareness of events in news bulletins
Document Visualisation - Stages

• Representation of results
  - form high-dimensional vector (one for each word, ~10000+)
  - cluster documents based on vector similarity (e.g. Nearest-Neighbour)

• Visualisation of clustered results
  - projection to lower dimensional space
  - 3D “galaxy” / 2D “theme-scape” / 1D “theme-river”

Query

“keywords” from user specification
comparison to sample “reference” document
2D and 3D projections of documents

3D Visualisation of 567,000 cancer literature abstracts.

Articles in a collection of news items (2D).

Pacific Northwest National Laboratory.
Multidimensional Scaling

A dimensionality reduction scheme useful for information visualization

Visualizing the information contained in a distance matrix

Computing the 2D coordinates of the projected samples such that the following stress is minimized

\[
Stress_D(x_1, x_2, \ldots, x_N) = \left( \sum_{i \neq j=1,\ldots,N} (d_{ij} - \|x_i - x_j\|^2)^2 \right)^{1/2}
\]

- distance between sample \(i\) and \(j\)
- coordinates of \(i\) and \(j\)
1D visualisation of news articles

A ‘Theme River’ shows the relative importance of themes over the course of a year from press articles.

Pacific Northwest National Laboratory.
Wordle

http://www.wordle.net/create

Produces a word cloud from a document
Document Querying

• Keyword search is problematic
  - ambiguity
  - ~7-18% of people describe same concept with same word (Barnard '91)

• Interested in
  - distribution of keywords in the document
  - related articles to the keyword entered

• Tile bar scheme (Hearst 1995)
  - display a list of documents with a tile bar
  - tile bar shows the occurrence of keywords in document
Columns represent paragraphs or pages in a document. Shade indicates relevance shown by word occurrence. Shows length and likely relevance. System allows interactivity by clicking on box.

- Visualisation - Use of document topology / colour-mapping / interaction
Example : Title Bar Query / Result

Query terms: DBMS (Database Systems)
Reliability

What roles do they play in retrieved documents?

Mainly about both DBMS & reliability
Mainly about DBMS, discusses reliability
Mainly about, say, banking, with a subtopic discussion on DBMS/Reliability
Mainly about something different
DocuBurst

- A radial, space-filling layout of hyponymy (IS-A relation)

```plaintext
- games → game
  - taken → take

- absolute, noun, 10
- chair, noun, 2
- moment, noun, 11
- game, noun, 30
- reality, noun, 3
- take, verb, 13
- represent, verb, 17

- game IS activity
  - chair IS furniture
```
Summary

• Information Visualisation
  – Univariante, bivariante, trivariante, multi-variate data
  – Relations visualized by lines, tree visualization
  – Document visualization
Reading

- Marti A. Hearst *TileBars: Visualization of Term Distribution Information in Full Text Information Access*


- Westin et al. ‘02, “Processing and visualization for diffusion tensor MRI”

- Watanabe et al., “Biclustering Multivariate Data for Correlated Subspace Mining”, PacVis 2015

- Wu et al. Spatially Efficient Design of Annotated Metro Maps, EuroVis 2013

- http://faculty.uoit.ca/collins/research/docuburst/index.html