SolarMap: Multifaceted Visual Analytics for Topic Exploration

Nan Cao, David Gotz, Jimeng Sun, Yu-Ru Lin, Huamin Qu
Abstract—Clustering as a fundamental data analysis technique has been widely used in many analytic applications. However, it is often difficult for users to understand and evaluate multidimensional clustering results, especially the quality of clusters and their semantics. For large and complex data, high-level statistical information about the clusters is often needed for users to evaluate cluster quality while a detailed display of multidimensional attributes of the data is necessary to understand the meaning of clusters. In this paper, we introduce DICON, an icon-based cluster visualization that embeds statistical information into a multi-attribute display to facilitate cluster interpretation, evaluation, and comparison. We design a treemap-like icon to represent a multidimensional cluster and the quality of the cluster can be conveniently evaluated with the embedded statistical information. We further develop a novel layout algorithm which can generate similar icons for similar clusters, making comparisons of clusters easier. User interaction and clutter reduction are integrated into the system to help users more effectively analyze and refine clustering results for large datasets. We demonstrate the power of DICON through a user study and a case study in the healthcare domain. Our evaluation shows the benefits of the technique, especially in support of complex multidimensional cluster analysis.

Index Terms—Visual Analysis, Clustering, Information Visualization.
Congenital heart disease

Overview
Congenital heart disease refers to a problem with the heart’s structure and function due to abnormal heart development before birth. Congenital means present at birth.

Symptoms
Symptoms depend on the type of heart defect.

Treatment
Treatment depends on the type of heart defect and may include medications and surgery.

Causes
Congenital heart disease occurs when a baby is born with a defect caused by the heart not developing properly. These defects must be treated.

While congenital heart disease may be immediately obvious right after birth, in some cases it cannot be seen for months or years. Other heart defects may never cause any problems and lead to a normal life span.

Tests & Diagnosis
Testing is important to determine the type of heart defect and how severe it is.

Prognosis
Outcomes vary depending on the type and severity of the defect.

Prevention
Some cases of congenital heart disease are preventable.

Complications
Some congenital heart diseases can be treated with medication alone, while others require one or more surgeries. The risk of death from congenital heart disease surgery has dropped from about 30% in the 1970s to less than 5% in most cases today.

Congenital heart disease is often divided into two types: cyanotic (blue discoloration caused by a relative lack of oxygen) and non-cyanotic. The following lists cover the most common congenital heart defects.
(Q1) How to model the document contents into multifaceted relation data?

(Q2) How to intuitively visualize multifaceted document contents and their relations?

(Q3) How to find interest patterns over multifaceted documents?
How to model the document contents into multifaceted relation data?

How to intuitively visualize multifaceted document contents and their relations?

How to find interest patterns over multifaceted documents?

How to explore the multifaceted documents to uncover their innate relations and topic evolutions?
Solution

- Goal:
  - Visualize the **multifaceted relations (both static and dynamic)** over documents

- Approach:
  - Multifaceted entity-relational data model
  - Intuitive visual encoding and automatic layout
  - Users’ interests driven interaction for pattern detection
Demo

1995
Key Challenges

(Q1) How to model the document contents into multifaceted relation data?

(Q2) How to intuitively visualize multifaceted document contents and their relations?

(Q3) How to find interest patterns over multifaceted documents?
(Q1) How to model the document contents into multifaceted relational data?

Document set → facet segmentation → entity extraction → entity set → multifaceted entity relational data model

**Facets:**
- **disease**
- **symptom**
- **treatment**

**Entity Extraction:**
- **type 1 diabetes**
- **type 2 diabetes**
- **thirst**
- **blurred vision**
- **take medications**
- **blood sugar control**

**Internal relations:**
- Type-1-Diabetes
- Type-2-Diabetes

**External relations:**
- Increased-thirst
- Blurred-vision
- Blood-suger-control
Key Challenges

(Q1) How to model the document contents into multifaceted relation data?

(Q2) How to intuitively visualize multifaceted document contents and their relations?

(Q3) How to find interest patterns over multifaceted documents?
(Q2) How to visualize multifaceted document contents and their relations?
How to represent the data model: Encoding

- Design Principles
  - Focus + Context
  - Content + Relations
  - Rich interaction
How to represent the data model: Encoding

- Visualize the **primary facet** in the center as topic contours.

![Diagram of topic clusters and keywords](image)
How to represent the data model: Encoding

- Visualize the **primary facet** in the center as topic contours
- Visualize the **context facet** around the center

Diagram:
- Topic clusters
- Type-2-Diabetes
- Type-1-Diabetes
-Keyword clusters
- Thirst
- Vision
- Symptoms
- Treatment
- Others

Keywords: 
- diabetes
- condition
- symptoms
How to represent the data model: Encoding

- Visualize the **primary facet** in the center as topic contours

- Visualize the **context facet** around the center

- Connect context entities with the primary entities using **links**

- Use colors to represent clusters

- Use distance to represent similarities (between clusters, nodes)
How to represent the data model: Encoding

- Visualize the **primary facet** in the center as topic contours
- Visualize the **context facet** around the center
- Connect context entities with the primary entities using **links**
- Use **colors** to represent cluster

**keywords clusters**

**topic clusters**

- symptom
- treatment
- other facet

**Type-1-Diabetes**

- increased thirst
- blurred vision

**Type-2-Diabetes**

- hungry

**keywords**
How to represent the data model: Encoding

- Visualize the **primary facet** in the center as topic contours
- Visualize the **context facet** around the center
- Connect context entities with the primary entities using **links**
- Use **colors** to represent cluster
- Use **distance** to represent similarities (between clusters, nodes)
(Q2) How to visualize multifaceted document contents and their relations?

data model

encoding

layout
Layout

- topic cluster layout
- keyword cluster layout
- Temporal sequence layout

Stabilized Layout

Based on the hidden internal relations of primary facet
Keep users’ mental map while data changed

\[
\min \left( \sum_{i<j} \frac{1}{d_{ij}} \left( \|X_i - X_j\| - d_{ij} \right)^2 + \sum_{i<j} \|X_i - \text{pre}(X_i)\|^2 \right)
\]

Cluster Together More smoothly
**Layout**

- **Topic cluster layout**
- **Keyword cluster layout**
- **Temporal sequence layout**

- **Topic cluster center detection**
- **Keyword wedge reordering**
- **Force directed cluster alignment**
Graph Sequence:
\[ GS = \{ G_1, G_2, \cdots, G_T \} \]

Segmentation Error:
\[ sse(S_i) = \sum_{s_i \leq j \leq e_i} ||G_j - S_i|| \]

Dynamic Programming:
\[ F(T, k) = \min_{t < T} (F(t, k - 1) + sse(S_k)) \]
Key Challenges

(Q1) How to model the document contents into multifaceted relation data?

(Q2) How to visualize multifaceted information to reveal both global and local patterns?

(Q3) How to find interest patterns over multifaceted documents?
Case Study 1: Healthcare
Case Study 2: DBLP

Year 1996
Case Study 2: DBLP
Evaluation: User Study

ContexTour (Lin et. al SDM 2010)  FacetAtlas (Cao et. al InfoVis 2010)

Baseline Systems
T1: Identify all clusters of diseases that match the query term “diabetes”. (cluster convey)

T2: Identify the top 3 symptoms for a specified disease cluster. (cluster interpretation)

T3: Identify the top 3 symptoms shared between two specified disease clusters. (cluster comparison across specific facets)

Significant improvement on cluster interpretation
Evaluation: Objective

(a) Comparison

(b) Usability of SolarMap
SolarMap: Multifaceted Visual Analytics for Topic Exploration

Nan Cao, David Gotz, Jimeng Sun, Yu-Ru Lin, Huamin Qu
The DBLP computer science bibliography

- Huge in amount
  - Over 1GB paper indices

- Dynamic in nature
  - Change of research communities and topics

- Multiple information facet
  - Authors
  - Conferences
  - Topics
(Q1) How to model the complex data?

(Q2) How to encode the data model in an intuitive way?

(Q3) What patterns can we find?
How to model the multifaceted document contents into multifaceted relation?

How to intuitively visualize multifaceted document contents and their relations?

How to monitor the source to uncover information dynamic patterns?

How to visualize the relations of multifaceted document contents?