FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Visual Analytics of Large Homogeneous Data

Categorical, Set-typed, and Classification Data

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der technischen Wissenschaften

eingereicht von

Bilal Alsallakh
Matrikelnummer 0627567

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung:
Ao. Univ.-Prof. Dr. Mag. Silvia Miksch
Prof. Dipl.-Ing. Dr. Helwig Hauser

Diese Dissertation haben begutachtet:

(Ao. Univ.-Prof. Dr. Mag. Silvia Miksch)	(Professor in Visualization Dipl.-Ing. Dr. Helwig Hauser)	(Professor John Stasko)

Wien, 05.08.2014

(Bilal Alsallakh)

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Visual Analytics of Large Homogeneous Data

Categorical, Set-typed, and Classification Data

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

PhD in Computer Science

by

Bilal Alsallakh
Registration Number 0627567

to the Faculty of Informatics
at the Vienna University of Technology

Advisors:
Ao. Univ.-Prof. Dr. Mag. Silvia Miksch
Prof. Dipl.-Ing. Dr. Helwig Hauser

The dissertation has been reviewed by:

(Ao. Univ.-Prof. Dr. Mag. Silvia Miksch)	(Professor in Visualization Dipl.-Ing. Dr. Helwig Hauser)	(Professor John Stasko)

Vienna, 05.08.2014

(Bilal Alsallakh)

Erklärung zur Verfassung der Arbeit

Bilal Alsallakh
Speisinger Strasse 57, 1130 Vienna

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Verfasser)

Acknowledgements

I acknowledge the contributions of the following people to this thesis:

- **Lilas Al Sallakh, my sister,** proofread this thesis thoroughly and pointed me to unclarities and writing issues.
- **Silvia Miksch** introduced me to the field of Visual Analytics, and how to conduct research in this field. She taught me the importance of task-driven and user-centered design, which had major influence on my research. She was open and supportive to explore new research directions that lead to this thesis, and dedicated a lot of time to discuss and refine my ideas and guide them into fruitful directions.
- **Helwig Hauser** gave crucial feedback on my ideas and proposed research directions that improved the applicability of my ideas to set visualization and classification data. He also referred me to inspiring related work and helped in structuring this thesis.
- **Eduard Gröller** introduced me to the field of visualization, and supervised my Master's thesis, where the first ideas behind this work started. He taught me scientific writing by providing grass-roots feedback on my Master's thesis and co-authored articles.
- **Wolfgang Aigner** helped me to refine my visualization and interaction design and to base it on solid theory. He also helped me to realize potential applications of my ideas.
- **Peter Filzmoser** gave statistical feedback on my proposals and provided useful datasets. He also encouraged me to applying my ideas to classification problems.
- **Chris Weaver** gave thoughtful feedback on my ideas on visualizing categorical associations and relations.
- **Aidan Slingsby** gave thoughtful feedback on my initial visual design and proposed the use of aggregation to simplify this design.
- **Simone Kriglstein, Margit Pohl, and Florian Scholz** gave critical feedback on my initial visual design which made me reconsider and refine many design decisions. Florian also influenced the user interface design by proposing to divide the space between overview and detail views and to show them side by side.

- **Markus Bögl** explained statistical concepts to me, and gave feedback on parts of this thesis. He also translated the abstract into German, together with Birgit.
- **Mohamad Rabbath** exchanged thoughtful discussions with me on measuring usability and performing evaluation. He also gave feedback on the introductory part of this thesis.
- **Stephan Hoffmann** gave feedback on parts of this thesis.
- **The IEG group at Vienna University of Technology** were part of a professional and nice working environment and gave constructive feedback on my test talks and presentations.
- **The anonymous reviewers** of the publications in Part II raised many important points that helped me refine my ideas and better explain my manuscript.

I also acknowledge the support of Mondi Austria Privatstiftung in financing my master studies in Vienna. It was during this period that the first ideas behind this thesis were developed.

Mrs. Fleury-Sauer was also very supportive throughout the course of this PhD. She would visit me in the lab in the last weeks of March, bring warm meals, and encourage me to work.

Finally, I dedicate this thesis to my father, M. Zafer. He dedicated a lot of his time to teach me math and physics and pushed me to read scientific books. Though he wished I study medicine, he supported my decision to study computer science with the condition that I earn a doctorate in this field instead. This was a constant motivating factor to pursue this thesis.

Related Publications

This thesis is based on the following publications (see Part II of the thesis):

- **Paper A:**

  Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and Helwig Hauser
  Radial Sets: Interactive Visual Analysis of Large Overlapping Sets
  IEEE Transactions on Visualization and Computer Graphics (Proceedings of IEEE InfoVis), 19(12):2496-2505, 2013.
  Supplementary video: `http://radialsets.org/`

- **Paper B:**

  Bilal Alsallakh, Allan Hanbury, Helwig Hauser, Silvia Miksch, and Andreas Rauber
  Visual Methods for Analyzing Probabilistic Classification Data
  IEEE Transactions on Visualization and Computer Graphics (Proceedings of IEEE VAST), 20(12), 2014, to appear.
  Supplementary video: `http://www.cvast.tuwien.ac.at/ConfusionAnalysis`

- **Paper C:**

  Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and M. Eduard Gröller
  Reinventing the Contingency Wheel: Scalable Visual Analytics of Large Categorical Data
  IEEE Transactions on Visualization and Computer Graphics (Proceedings of IEEE VAST), 18(12):2849-2858, 2012.
  Supplementary video: `http://www.cvast.tuwien.ac.at/wheel`

In addition, the following publications are related to the topic of this thesis:

- The early version of the proposed visual metaphor:
  Bilal Alsallakh, Eduard Gröller, Silvia Miksch, and Martin Suntinger
  Contingency Wheel: Visual Analysis of Large Contingency Tables
  Proceedings of the International Workshop on Visual Analytics (EuroVA), pp. 53-56, 2011, Eurographics.

- Evaluation of the early version of the proposed visual metaphor:
  Margit Pohl, Florian Scholz, Simone Kriglstein, Bilal Alsallakh, and Silvia Miksch

Evaluating the Dot-Based Contingency Wheel: Results from a Usability and Utility Study
Proceedings of HCI International - Lecture Notes in Computer Science (LNCS) on *Human Interface and the Management of Information, pp. 76-86, 2012, Eurographics.*

- A state of the art report on visualizing set-typed data:
  Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch, and Peter Rodgers
  Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges"
  Proceedings of the Eurographics Conference on Visualization (EuroVis) - State of The Art Reports, pp. 1-21, 2014,Eurographics.

- An application of the proposed set visualization in Information Retrieval:
  Bilal Alsallakh, Silvia Miksch, and Andreas Rauber
  Towards a Visualization of Multi-faceted Search Results"
  Proceedings of the DL2014 Workshop on Knowledge Maps and Information Retrieval (KMIR), the ACM/IEEE Joint Conference on Digital Libraries, 2014.

Finally, the following is a list of other publications I worked on during my PhD study and as part of my work in the CVAST project. These publications are not closely related to the thesis topic.

- Bilal Alsallakh, Markus Bögl, Theresia Gschwandtner, Silvia Miksch, Bilal Esmael, Arghad Arnaout, Gerhard Thonhauser, and Philipp Zöllner
  A Visual Analytics Approach to Segmenting and Labeling Multivariate Time Series Data"
  Proceedings of the International Workshop on Visual Analytics (EuroVA), 2014, Eurographics.

- Bilal Alsallakh, Peter Bodesinsky, Silvia Miksch, and Dorna Nasseri
  Visualizing Arrays in the Eclipse Java IDE
  European Conference on Software Maintenance and Reengineering, 2012, pp. 541-544, IEEE.

- Bilal Alsallakh, Peter Bodesinsky, Alexander Gruber, and Silvia Miksch
  Visual Tracing for the Eclipse Java Debugger
  European Conference on Software Maintenance and Reengineering, 2012, pp. 545-548, IEEE.

Abstract

A multidimensional data set is homogeneous when the dimensions have the same nature. For instance, these dimensions can represent the probabilities for a sample to belong to different classes, or item memberships of multiple sets. Such data appear very often in different domains to describe how a relatively large number of items are related to a relatively small number of classes or categories. For examples, a homogeneous data set might record which genes (rows) appear in which individuals (columns), or how many times books (rows) are sold in different countries (columns). Analyzing these relations reveals several patterns in the data such as genes that are observed frequently or never together, or books that sell mostly in a specific country. Both automated methods and visualization have been applied to analyze homogeneous data. However, state-of-the-art visualization techniques are lacking either in scalability with the number of data points or in addressing the specific nature of different classes of homogeneous data, and the tasks associated with them.

In this dissertation, I propose novel *visual metaphor* and *interactive exploration environment* for analyzing large homogeneous data. The proposed *wheel* metaphor allows analyzing and selecting the data points based on their relations with the different dimensions. Moreover, it emphasizes the dimensions and the relations between them as the central part of the visualization, and allows analyzing these relations based on the data points defining them. The proposed *interactive exploration environment* allows analyzing different aspects of the data at multiple levels of detail. I illustrate how the proposed approach can be applied to analyze three classes of homogeneous data: *set-typed data*, *probabilistic classification data*, and *categorical data*. Each class has its own characteristics that imply specific requirements and tasks. These different tasks are supported by the proposed approach, thanks to its flexibility and extensibility.

I demonstrate the applicability of my approach by means of usage scenarios and case studies with various datasets from multiple domains. Also, both user studies and interviews with domain experts were conducted to assess the utility of the proposed methods. The major advantages of the proposed visual metaphor is its scalability in the number of data points, thanks to dedicated aggregation methods for homogeneous data, and to the rich sets of interactions it supports to select the data based on a variety of criteria. The major disadvantages are the complexity of the visual metaphor that requires sufficient user training, the limited scalability in the number of dimensions, and the low sensitivity to small differences in the data being analyzed. Nevertheless, the wheel metaphor is suited to gain an overview of large homogeneous data, with complementary analytical methods, interactions, and coordinated views being used to cope with the limitations. As a result, novel analysis possibilities and insights in the data are possible, beyond state-of-the-art techniques.

Kurzfassung

Homogene multivariate Daten umfassen eine Vielzahl an Variablen mit ähnlichem Verhalten/ ähnlicher Struktur. Diese Variablen können Unterschiedliches repräsentieren – zum Beispiel die Wahrscheinlichkeit, mit der ein Element in eine bestimmte Gruppe gehört, oder die Zuordnung eines Elementes zu einer Reihe von Mengen. In vielen Anwendungsgebieten werden solche Daten genutzt, um die Zugehörigkeit von einer relativ großen Menge an Elementen zu einer relativ kleinen Anzahl an Gruppen oder Kategorien zu beschreiben. Eine homogene Tabelle zeigt beispielsweise, welche Gene (Zeilen) in welchem Individuum (Spalten) vorkommen, oder wie oft ein Buch (Zeilen) in verschiedenen Ländern (Spalten) verkauft wurde. Die Analyse solcher Zusammenhänge ermöglicht es, Muster in den Daten zu erkennen – etwa Gene, die oft oder nie zusammen vorkommen, oder Bücher, die hauptsächlich in bestimmten Ländern verkauft werden. Für die Untersuchung derartiger Muster in großen homogenen Datenmengen wurden bereits automatisierte Methoden und Visualisierungen angewandt. Allerdings mangelt es selbst bei der Verwendung neuester Visualisierungstechniken an der Skalierbarkeit in Bezug auf die Anzahl von Elementen, und an der fehlenden Miteinbeziehung der speziellen Eigenschaften, die verschiedene Gruppen homogener Daten, bezogen auf die konkreten Aufgabenstellungen, haben.

In dieser Dissertation stelle ich neue visuelle Metaphern und interaktive Explorationsumgebungen für die Analyse großer homogener Daten vor. Die vorgeschlagene Rad-Metapher ermöglicht es, basierend auf den Zusammenhängen mit anderen Spalten, Elemente auszuwählen und zu untersuchen. Darüber hinaus liegt das Hauptaugenmerk der Visualisierung auf den Spaltenvariablen und den Relationen zwischen den Spalten. Dieser Fokus ermöglicht die Analyse dieser Beziehungen basierend auf den Zeileneinträgen, die diese Relationen definieren. Die interaktive Explorationsumgebung erlaubt es, verschiedene Aspekte der Daten und der Element-Attribute in verschiedenen Detailgraden zu betrachten. Ich veranschauliche meinen Ansatz mit drei unterschiedlichen Arten von homogenen Daten: mengenartige Daten, wahrscheinlichkeitstheoretische Klassifikationsdaten, und kategorische Daten. Jede dieser drei Gruppen weist bestimmte Charakteristika in den Daten auf, wie etwa spezielle Anforderungen und Aufgaben. Damit zeige ich, dass die visuelle Metapher ausreichend flexibel und erweiterbar ist, um diese Aufgaben skalierbar zu lösen.

Ich belege die Anwendbarkeit meines Ansatzes anhand von Usage-Szenarien, Insight-Studien und Fallstudien mit unterschiedlichen Daten aus mehreren Domänen. Zur Beurteilung der Brauchbarkeit der vorgestellten Methoden wurden Benutzerstudien und Interviews mit Experten durchgeführt. Die größten Vorteile der visuellen Metapher sind die Skalierbarkeit in Bezug auf die

Anzahl der Elemente anhand von geeigneten Aggregationsmethoden für homogene Daten, sowie die zahlreichen Interaktionsmöglichkeiten, um die Auswahl der Daten basierend auf einer Vielzahl von Kriterien zu unterstützen. Nachteile zeigen sich in der Komplexität der visuellen Metapher, welche es für den Benutzer notwendig macht, diese ausreichend zu erlernen, in der limitierten Skalierbarkeit in Bezug auf die Anzahl der Spalten und in der niedrigen Sensitivität, kleine Unterschiede in den Relationen zu analysieren. Dennoch ist die Rad-Metapher geeignet, die Limitierungen mit komplementären analytischen Methoden, Interaktionen und koordinierten Ansichten zu überbrücken und damit einen Überblick über große homogene Daten zu erlangen. Als Ergebnis entstehen neuartige Analysemöglichkeiten sowie neuartige Erkenntnisse in den Daten, und zwar über den aktuellen Stand der Technik hinaus.

Contents

List of Figures

Part I

Overview

CHAPTER 1

Introduction

1.1 Motivation

Background

Information Visualization (*InfoVis*) and Visual Analytics (*VA*) have emerged as new data analysis paradigms that complement automated techniqus with visualization. *InfoVis* has been defined as "the use of computer-supported, interactive, visual representations of *abstract data* to amplify cognition" [24, p. 7]. Such data have usually no inherent spatial structures that characterize scientific data such as blood flow or air density. Abstract data often represent non-physical entities and concepts such as friendship relations between people, tags on documents, or the hierarchy of an organization. The data can originate from a variety of domains, and hence can vary strongly in form (e.g., tabular, hierarchical, relational, etc.), semantics, and dimensionality. *VA* has been defined as "the science of analytical reasoning, facilitated by interactive visual interface" [151, p. 4].

A large number of *InfoVis* and *VA* techniques have been developed over the past two decades to support the analysis of different types of data such as trees [128], graphs [161], and time-oriented data [105]. Besides reference collections of such techniques, a lot of effort has been made to develop data type and task taxonomies which help the systematic development, understanding, and evaluation of these techniques [20, 139]. Understanding the characteristics of different data types and identifying common analytical tasks associated with them enable better analysis of the design space of visualization techniques for these types, and fosters the re-usability and the wide applicability of these techniques. Shneiderman's *Task by Data Type Taxonomy* [139] was the first and most foundational work on developing InfoVis data type and task taxonomies. With respect to data type, Shneiderman identified seven generic data types that have different characteristics: "one-, two-, three-dimensional data, temporal and multi-dimensional data, and tree and network data". Besides these types, he identified seven generic tasks that can be applied to these datatypes: "one-, two-, three-dimensional data, temporal and multi-dimensional data, and tree and network data".

In his taxonomy, Shneiderman illustrated how relational and statistical databases can be conveniently manipulated as multidimensional data in which items with m attributes become points in an m-dimensional space. These attributes might represent real-world concepts having different nature such as length and weight, making the data potentially heterogeneous. Heterogeneous data encompass multiple attributes that can potentially have different forms (e.g. numerical /categorical / Boolean attributes), different measurement units (kg, meter, USD), or different ranges of possible values.

Homogeneous Data: A Special Type of Data

In some applications, all the attributes of a multidimensional dataset have the same nature and value ranges. As an example, the data can represent the probabilities computed by a classifier for a set of n handwritten letters to belong to each of the 26 letter classes. In this case, all the attributes have the same nature: the probability for a sample to belong to a class. Also, the values of these attributes can be compared directly as they belong to the same value range $[0, 1]$. Moreover, the attribute values for one sample can be aggregated meaningfully. For example, the sum of these values is equal to 1. Also, the maximum value indicates the class with the highest probability, which is selected as the predicted class for the given sample by the classifier.

In this thesis, I refer to multidimensional datasets whose attributes have the same nature and value ranges as *homogeneous data*. I argue that these special characteristics provide new possibilities to analyze and visualize such data, beyond the possibilities available for heterogeneous data (Chap. 3). I demonstrate these possibilities with three classes of homogeneous data: (1) set-typed data that encompass element-set memberships, (2) classification data that encompass sample-class probabilities, and (3) two-way contingency tables that describe the relation between two categorical variables. Each of these classes has been treated separately in the literature, with several methods proposed to visualize data of this class. These methods have limitations either in terms of scalability, or in the level of detail they reveal in the data. I demonstrate how the similar structure of the above-mentioned classes allows specifying many tasks with the data in similar ways and developing common visualization and interaction methods to solve these tasks.

Besides homogeneous attributes, many datasets contain additional heterogeneous attributes that describe the row entities. For example, in addition to class probabilities, the data about classification samples such as handwritten letters might encompass classification features used by the classifier to compute these probabilities. These attributes often determine the values of homogeneous data or influence certain tendencies in them. I demonstrate how the methods I propose allow analyzing these heterogeneous attributes in relation to the homogeneous data.

1.2 Problem Statement

As proposed by Miksch and Aigner [104], I define the *VA* problem addressed in this work by means of the *data* being analyzed, the *users* who will use my methods, and the *tasks* these users need to solve with the data. Additionally, I state the hypothesis behind my research and the corresponding research questions.

Data Features			Classification Results		Class Probabilities						
F1	...	Fp	Actual	Predicted	drawing	graph	math	table	flowchart	...	character
5		62	flow_chart	flow_chart	0.001	0.000	0.000	0.000	0.999		0.000
23		11	graph	drawing	0.502	0.498	0.000	0.000	0.000		0.000
1		12	drawing	drawing	0.051	0.947	0.000	0.000	0.000		0.002
343		42	graph	graph	0.000	1.000	0.000	0.000	0.000		0.000
54		45	drawing	drawing	1.000	0.000	0.000	0.000	0.000		0.000
⋮			⋮		⋮		⋮		⋮		
33		12	drawing	graph	0.228	0.772	0.000	0.000	0.000		0.000

Figure 1.1: A classification dataset containing three groups of attributes of the classification samples: (1) heterogeneous data features used for classification, (2) classification results (actual and predicted class names), (3) *homogeneous attributes* representing the probabilities computed for the samples to belong to each of the classes.

1.2.1 Data

This work is dedicated to analyze homogeneous data, as a special type of multidimensional data. I define homogeneous data as a dataset of n entities E described by a set of m homogeneous numerical attributes $A_1 \dots A_m$ having the same nature and value ranges. The entities E can have additional p heterogeneous attributes $F_1 \dots F_p$ as well.

In particular, I address homogeneous data having a relatively small number of homogeneous attributes m (in the order of tens), but a relatively large number of entities n (in the order of tens of thousands). I identify three classes of data that can be modeled as homogeneous data:

- **Element-Set memberships (set-typed data):** given m sets $S_1 \dots S_m$, the membership of the n entities E to these sets can be represented by m homogeneous attributes $A_1 \dots A_m$. The values of each binary attribute A_j denote the membership of the elements to the respective set S_j, with $A_j(e_i \in E) = 1 \iff e_i \in S_j$ (Fig. 1.2a).

  As an example, the entities E can represent survey participants and the sets can represent m possible answers to a multiple-choice question. The attributes $A_1 \dots A_m$ denote which participant selected which answer(s) to the question. The participants might further have other heterogeneous attributes $F_1 \dots F_p$ such as age, sex, or political affiliation.

  I refer to this class data as **set-typed data** throughout this thesis.

- **Probabilistic classification data:** Probabilistic classifiers compute the probability $p_{ij} \in [0, 1]$ that a sample $e_i \in E$ belongs to one of m classes $C_1 \dots C_m$. These probabilities can be represented by m homogenous attributes $A_1 \dots A_m$ where $A_j(e_i) = p_{ij}$. The attribute values for item e_i sum up to $\sum A_j(e_i) = 1$. In addition, the samples can have a number of heterogeneous data features $F_1 \dots F_p$ used for the classification (Fig. 1.2b).

  As an example, the items E can be a large set of images that contain handwritten digits. The classes $C_1 \dots C_m$ represent the digits. The attribute value $A_j(e_j)$ indicates the probability that image e_i represents digit A_j, as computed by the classifier. The heterogeneous attributes $F_1 \dots F_p$ represent classification features extracted from the images.

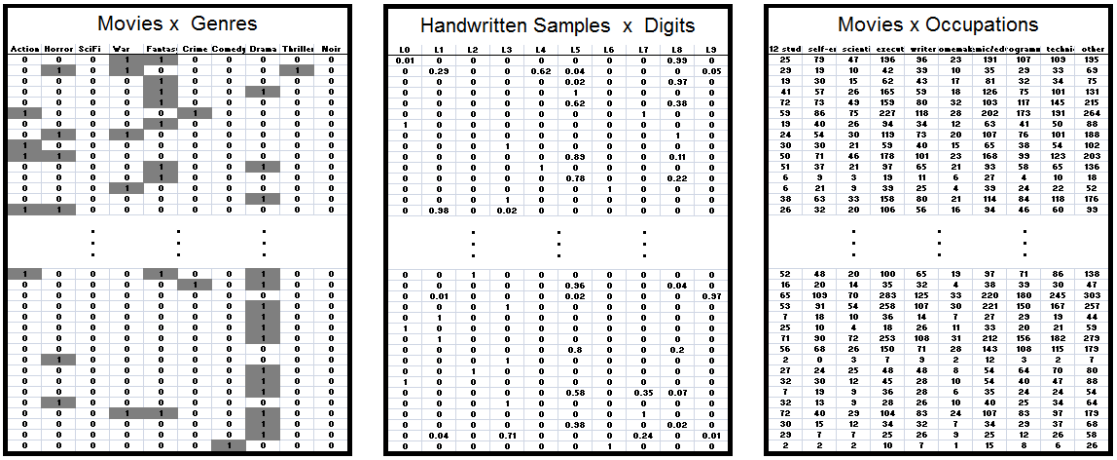


Figure 1.2: Three examples of homogeneous data having similar dimensionality: (a) sample-class probabilities, (b) element-set memberships, (c) a contingency table.

- **Contingency tables / Categorical data:** An $n \times m$ contingency table records the frequency of observations $f_{ij} \in \mathbb{N}$ for each combination of categories $(e_i, c_j) \in E \times C$ of two categorical variables E and C. In a skewed contingency table with $m << n$, the variable C has few categories while the row variable E has a large number of categories. This table can be modeled as m homogenous attributes $A_1 \ldots A_m$ representing the categories of variable C where $A_j(e_i) = f_{ij}$ (Fig. 1.2c).

  As an example, E can be a large set of books and C a set of countries, with f_{ij} representing the purchases of book $e_i \in E$ in country $c_j \in C$. In addition, these books can have a set of heterogeneous attributes $F_1 \ldots F_p$ such as release date, author(s) and publisher(s).

1.2.2 Users

The techniques I propose can be potentially applied to data of the above-listed classes, in a generic way. When applied to a specific application domain, these techniques require users who are experts in this domain and understand the data and their characteristics. Domain expertise is important to (1) determine if and how the proposed methods are applicable to the particular domain, (2) steer computational analysis, e.g., to choose appropriate algorithms, measures, and parameters for a specific dataset, and (3) interpret the visualization correctly, e.g., to relate the visual findings to the data and the problem domain, and to draw correct conclusions about them.

As an example, I assume that users who aim to analyze sample-class probabilities are machine-learning experts who are aware of the classification algorithm and data feature used to compute these probabilities. Similarly, users who want to analyze relation between genes and individuals are assumed to be biologists who can interpret the data.

1.2.3 Tasks

It is important to understand which particular tasks need to be carried out by the users during the analysis process, in order to design effective *VA* solutions that address these tasks [104]. In his *Task by Data Type Taxonomy* [139], Shneiderman identified seven abstract tasks that arise when using *InfoVis* applications: overview, zoom, filter, details-on-demand, relate, history, and extract. Schultz et al. [130] propose a *design space* of visualization tasks according to the following dimensions: goal, means, data characteristics, data target and cardinality, the order of tasks, and the (type of) user. They apply this space to recommend suited visualization for a given visualization task. Brehmer et al. [20] propose a *multi-level typology* of abstract visualization tasks. Their topology aims to connected low-level and high-level task taxonomies by describing complex tasks as a sequence of interdependent simpler ones. They organize this typology on three dimensions: *why* the task is performed, *how* the task is performed, and *what* are the task's inputs and outputs.

For each of the three classes of data listed in the previous section, there are specific task requirements. For example, typical tasks with set-typed data involve analyzing how the elements belong to the sets and how the sets intersect. On the other hand, Analyzing classification data involves finding which samples have high or low probabilities to belong to a certain class, and comparing confusions between the classes. Analyzing contingency data, also involves finding categories that appear more often together than other categories.

A comprehensive list of tasks related to set-typed data has been proposed recently [5]. It encompasses tasks related to the elements E, tasks related to the sets $A_1 \cdots A_m$, and tasks related to the attributes $F_1 \cdots F_p$. Chapter. A illustrates which of these set-related tasks are addressed in this thesis. Chapter. B discusses analysis tasks with classification data proposed in the literature, and demonstrates the ones addressed in this thesis. Chapter. C demonstrates statistical analysis tasks of large contingency tables extracted from categorical data.

The structural similarity between the classes allows formulating many of their tasks in a unified way. For example, analyzing data of any of these classes involve finding data points that are highly associated with a specific homogeneous dimension e.g. by belonging exclusively to the respective set or by having very high probability to belong to the respective class. This suggests developing common interactive solutions for these tasks as I show in Chapter 3.

1.2.4 Hypotheses and Research Questions

The two main hypothesis of the proposed research are:

- **H1:** *VA methods provide new insights in- and analysis possibilities of homogeneous data, that are difficult to gain or perform using automated data analysis methods.*
- **H2:** *The structural similarities between different classes of homogeneous data enable re-using* VA *solutions across these classes.*

Based on the research problem and hypotheses stated above, I investigated *VA* methods to address the following research questions about homogeneous data:

- **Q1:** *How to analyze the association relations between a large number of row entities E and the m homogeneous column categories represented by the attributes $A_1 \dots A_m$?*

  For example, a basic question about set-typed data is to which sets (column categories) an element (row entity) belongs. Likewise, analyzing classification results involves finding at which probability certain samples belong to certain classes. Finally, finding pairs of categories that frequently appear together is a typical task in analyzing contingency tables.

- **Q2:** *How to analyze the correlation between homogeneous attributes $A_1 \dots A_m$?*

  Besides being dependent on the application domain, the definition of attribute correlation highly depends on the narrow class of the homogeneous data. For example, two attributes that correspond to a highly overlapping pair of sets take the same values (0 or 1) for a significant number of row entities. Also, when analyzing classification results, it is desirable to find classes that are frequently confused for each other. In contingency table analysis, correlations are usually computed using statistical methods such as correspondence analysis.

- **Q3:** *How to analyze the influence (in both directions) of an additional attribute F_k of the row entities on the row-column associations (Q1) and attribute correlations (Q2)?*

  For instance, elements that belong to a certain set (or set combinations) might predominantly have a high or a specific value for a certain attributes. As an example, movies that belong exclusively to genre "Horror" receive lower ratings on average than those of other genres. Also, the handwritten digits frequently confused between two classes might be written mainly be a certain test subject. Finally, row entities that have certain values for F_k might have disproportionally high frequencies with certain column categories in a contingency table. For example, language and publisher are usually factors that impact book purchases in specific countries, e.g., German books usually have disproportionately large purchases in German-speaking countries.

1.3 Contributions

The research contributions of this thesis are twofold. On one hand, this thesis contributes a set of tools for analyzing and visualizing three specific classes of data, in particular:

1. **Novel *VA* methods for analyzing set-typed data:** This technique, called Radial Sets, provides an overview of how a large number of elements (in order of tens of thousands), belong to about 30 to 40 sets. As demonstrated in Chapter A, Radial Sets enable flexible analysis of the data and provide new insights into the data that is not possible in state-of-the-art set visualization techniques.

2. **Novel *VA* methods for analyzing probabilistic classification data:** These methods provide novel ways for analyzing classification results in relation with classification probabilities computed by a probabilistic classifier. They can handle tens of thousands of samples classified into about 20 classes. As demonstrated in Chapter B these methods reveal several patterns in the classification results that allow improving the classifier design.

3. **Novel *VA* methods for categorical data:** These methods allow analyzing associations between the categories of two categorical variables, with one of them having a large number of categories. The respective skewed contingency table (tens of thousands $\times$ 30) cannot be handled using previous techniques such as mosaic displays [58] or Parallel Sets [13], that are usually restricted to a total number of categories in the order of tens.

On the other hand, this thesis identifies the class of homogeneous data as a subclass of multidimensional data. It describes the characteristics of this sub-class and demonstrates the limitations of generic multidimensional data analysis and visualization techniques when applied to homogeneous data. Based on that, the thesis makes the following contribution:

1. **A unified view on different classes of homogeneous data** that is based on the structural similarities between them, and that takes into account the variations between them. This view enables a common VA framework for these different classes.

2. **Unified automated analysis of large homogeneous data** to reduce the data volume and extract the most important information and relationships in the data. The common abstraction of the data takes the specific nature of different data classes into account, by employing class-specific analytical methods.

3. **A unified visual metaphor for homogeneous data** based on their common abstraction. This metaphor provides overview of major relationships in the data by visualizing, and allows various interactions to brush the data for exploration in other views.

4. **A unified interactive exploration environment for homogeneous data** that incorporates class-specific views and allows exploring the data at multiple levels of abstraction.

5. **Scalable analysis of homogeneous data w.r.t the number of data points**, thanks to automated analysis that aggregates the data and to the multi-level exploration environment.

1.4 Structure of the Thesis

This thesis is divided into two parts:

- Part I (Overview) illustrates the context of the work and motivates the special class of homogeneous data as well as the research questions of this thesis. Chapter 2 provides a summary of most related state-of-the art methods for visualizing multidimensional data, and why they are insufficient to visualize homogeneous data. It also provides an overview of general data reduction approaches for multidimensional data, and how they can be applied in the context of homogeneous data. Chapter 3 introduces the VA approach proposed in this thesis, along with a comparison with existing multidimensional visualization methods mentioned above, in addition to a summary of the evaluation results as well as the scalability and perceptual limitations of the proposed visualization. Chapter 4 demonstrates four selected applications of the proposed approach in different domains. Chapter 5 reflects on the lessons learned and illustrates how the proposed approach addresses the research questions presented in Sect. 1.2.4.

- Part II (Papers) explains how the proposed *VA* approach is applied to each of the three classes of homogeneous data addressed in this thesis. It consists of three research papers published in the IEEE Transactions on Visualization and Computer Graphics:
  - **Paper A** is about visualizing element-set memberships (set-typed data).
  - **Paper B** is about analyzing probabilistic classification data.
  - **Paper C** is about visualizing categorical data and contingency tables.

  These papers provide details about the three classes of data and the concrete tasks associated with them. Each paper provides state-of-the-art methods designed specifically for visualizing the respective class of data. The paper also illustrates in detail how the proposed VA approach is applied to this class by (1) explaining the automated analysis methods employed, (2) showing how the visual metaphor is adjusted to fit the nature of the data, and (3) illustrating additional class-specific views that are developed to support the interactive exploration of the data. The three papers provide several usage scenarios for different datasets and summarize feedback by domain experts when the method is used to analyze data in a specific applications domain. The papers reflect also on encountered limitations and propose extensions for future work.

CHAPTER 2

State of the Art

As discussed in Chapter 1, homogeneous data are a special class of multidimensional data. Therefore, visualization techniques for multidimensional data can basically be applied to homogeneous data, however, without addressing or exploiting the characteristics of such data.

Section 2.1 surveys related multidimensional visualization techniques, and discusses their applicability to homogeneous data Section 2.2 provides an overview of data reduction methods for multidimensional data. Section 2.3 provides an overview of exploration environments that employ multiple views to explore the data at multiple levels of detail. Part II elaborates on state-of-the-art visualization techniques to visualize sets, classification data, and categorical data.

2.1 Visualizing Multidimensional Data

A large portion of *InfoVis* research focuses on visualizing multidimensional data. Many categorizations have been proposed to classify existing visualization techniques [24, 81, 162], for example, based on the graphical primitives used in the rendering [162]: points, lines, regions, or combinations thereof. Keim [81] classifies the techniques into five categories, according to the visual representations they use to encode the data values:

- **Geometric techniques** project the data in a geometric space using position or size as primary visual variables [17] to encode the data values. Examples for this are scatterplot matrices (SPLOMs) [56], parallel coordinates [69], and RadViz [68].
- **Pixel-oriented techniques** encode the data values in a compact display using color as a primary visual variable. Examples for this are heat maps and pixel bar charts [78].
- **Icon-based techniques** use icons that encode the data values for individual data items. Examples for this are stick figures [111] and color icons [91].
- **Hierarchical techniques** recursively divide the data space to show the values of multiple attributes. Example for this are dimensional stacking [90] and tree maps [138].
- **Graph-based techniques** use node-link diagrams to show relations between data items and dimensions. Examples for this are Ploceus [93] and PivotSlice [175].

The above categories are not exclusive, as some techniques combine ideas from multiple categories. The following sections overview of representative techniques from three of the above categories. The techniques are selected based on (1) their relatively high scalability with the number of data points (compared with other techniques) and (2) their generic applicability to multidimensional data from different domains. To illustrate the applicability of the following techniques to visualize homogenous data, I provide a screenshot when possible of how they visualize the example dataset depicted in Fig. 1.1. The dataset comprises probabilities computed by a probabilistic classifier for 1000 patent images to belong to 9 classes of patent images. The data was collected during the CLEF-IP 2011 classification evaluation campaign [113]. Sect. 3.6 compares the proposed *VA* approach with these techniques in terms of scalability and the new insights it reveals in homogeneous data.

2.1.1 Scatterplot Matrices (SPLOMs)

SPLOMs [56] use a matrix layout to show the relations between pairs of dimensions by means of scatterplot. Each scatterplot depicts the whole data points projected on the respective pair of dimensions. The scatterplots are synchronized: brushing certain data points in one plot highlights these points in the other plots.

Fig. 2.1 shows a SPLOM of the classification probabilities depicted in Fig. 1.1. Each scatterplot shows the relation between classification probabilities of the samples in the respective pair of classes. The majority of samples in Fig. 2.1 are located along the two axis in their scatterplots. These samples were determined by the classifier to possibly belong to one of the class, and not to the other class in the respective scatterplot. Fewer samples were assigned non-zero probability in both of the classes. These samples appear more often when the respective pair of classes are more likely to be confused for each other (e.g. “graphs”, “drawings”, and “flow chart”).

Depending on the complexity of the data, SPLOMs can fit up to 10-20 dimensions at a sufficient resolution. Also, since the plots are allocated small areas, they can fit only a relatively small number of data points (up to few hundreds). SPLOMs are usually used to depict the raw data values, without additional aggregated information about the data elements. This further limits their applicability to homogeneous data, when such aggregations are required. For example, if the homogeneous variables represent binary set memberships, the data is mapped to one of the four points $(0,0),(0,1),(1,0),(1,1)$, which provides limited insights into the data.

2.1.2 Parallel Coordinates

Proposed by Inselberg [69], parallel coordinates are one of the popular geometric techniques to visualize multidimensional data in one plot. The dimensions $A_1..A_m$ are represented as equidistant parallel axes. The data items $e_1..e_n \in E$ are represented as polylines. Each polyline connects the points on the parallel axes that correspond to a tuple in E.

Fig. 2.1 shows a parallel-coordinates plot of the classification probabilities depicted in Fig. 2.1. Each of the 9 axes represents a class and each of the 1000 polylines represents a sample. It is evident that samples having high probabilities in one class have low probability in other classes. This finding is already, as the probabilities sum up to 1. Also, few samples have high probabilities in class “gene seq” compared with class “graph”. this finding would have been possible

Figure 2.1: A scatterplot matrix of the class probabilities depicted in Fig. 1.1.

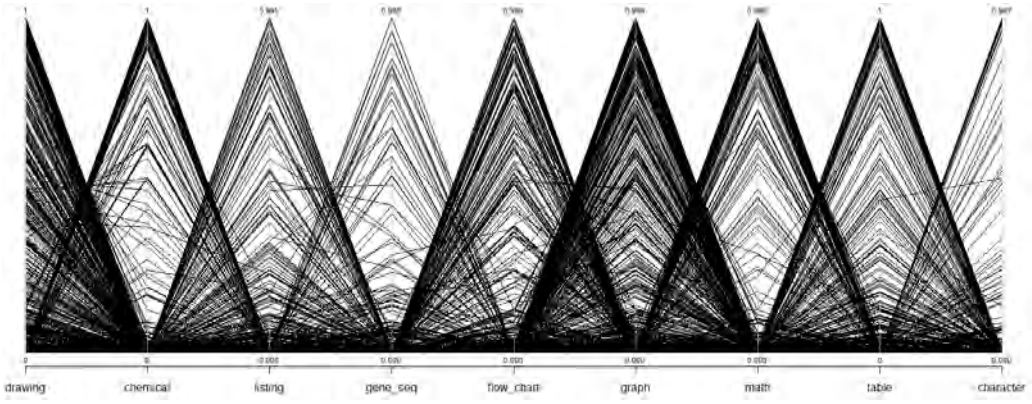


Figure 2.2: A parallel-coordinates plot of the class probabilities depicted in Fig. 1.1.

to observe using simple histogram charts of the sample probabilities in each class. Several interactions are possible, e.g., to change the axes ordering or scale and to filter the elements by brushing certain value ranges.

The main drawback of parallel-coordinates plots is clutter, which obscures several relations and patterns in the plot. Another problem is their sensitivity to axis ordering: certain patterns are revealed only when certain ordering is used. There has been a lot of work to reduce clutter and to compute suitable axis orders. Nevertheless, parallel-coordinates plots remain less suited to visualize homogeneous data, as they do not address the characteristics of such data that need specific aggregations to emphasize relevant patterns in the data.

2.1.3 Biplots

A biplot [50] is a 2D projection of multidimensional data, based on Principal Component Analysis (PCA) [75]. The dimensions are also projected on this plane, as vectors, to represent their contribution to the variance and their relations to the data values. The data points are projected on the 2D plane spanned by the first two principal components that capture the majority of the variance in the data. Fig. 2.3 shows the biplot of the classification probabilities depicted in Fig. 1.1. It reveals several patterns such as clusters of samples having similar projections. These samples have similar class probability profiles, compared with other samples. Furthermore, it is evident that classes "drawing" and "graph" have large contribution to the overall variance. Their samples appear in different forms in the data set, leading to varying probability profiles.

The main drawbacks of biplot are clutter and difficulty of interpretation. The principal components cannot be directly associated with the data dimensions that have clear interpretation. Also, overplotting restrict the usefulness of the plots to small datasets.

Several other projections of multidimensional data are possible, such as *Multidimensional Scaling* (MDS) [86]. Recently, Turkay et al. [154] propose a simplified projection on their *Dual Analysis* technique. This projection includes dimensions only, and allows users to select an interpretable 2D projection space. The authors showed how using common statistical measures to define this space, such as mean and variance, enables selecting attributes that exhibit interesting behavior and analyzing the respective values in the data space.

2.1.4 Table Lens

Instead of projecting multiple dimensions on the same plot, *table lens* [120] visualizes the values in each dimension individually, using a table layout. The table columns represent the dimensions, while rows represent the data items. The data values are represented as lines of proportional length and of 1-pixel thickness to fit a large number of items in one screen.

Fig. 2.4 shows a table lens view of the classification probabilities depicted in Fig. 1.1. Each column represents a class, with the bars showing sample probabilities to belong to this class. Columns have the same size and scale to enable direct comparison of lines. The table rows are sorted by their probabilities to belong to class "graph". This reveals the probability distribution of the samples in this class, in addition to how the other class probabilities correlate with class "graph". For example, it is evident that there are several samples which have relatively high probability value both in "drawing" and in "graph".

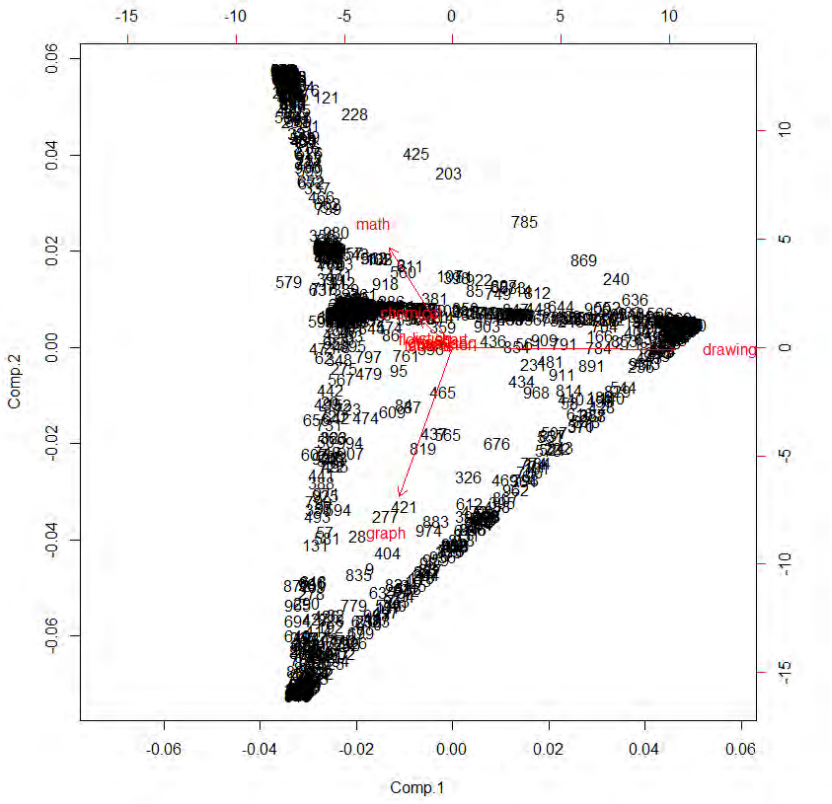


Figure 2.3: A biplot of the class probabilities depicted in Fig. 1.1. Labels in black represent the data points while labels in red represent the classes (the homogeneous dimensions).

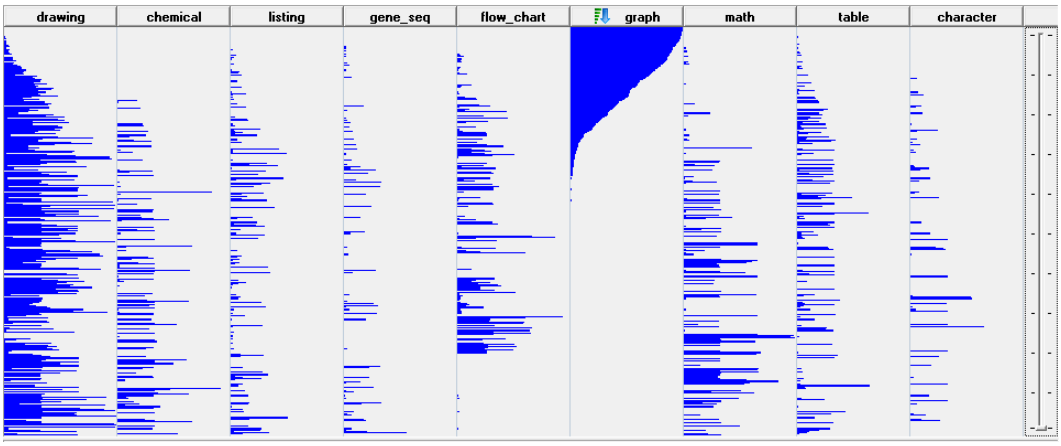


Figure 2.4: A *table lens* view of the classification probabilities depicted in Fig. 1.1.

Table lens combines features from both geometric and pixel-based techniques to provide a simple and compact overview of homogeneous data. However, the view is limited in the insights it reveals. The table can be sorted by one column at a time. This makes it hard to analyze the distribution of the other columns and to compare them against each other. Furthermore, the view suffers from over-plotting when showing more data items than pixels available. Finally, the metaphor restricts data aggregation possibilities that are needed to analyze certain homogeneous data such as set-typed data. This is because the values for a data point should be placed in the same row and cannot be aggregated to groups lying in different rows.

2.1.5 Heat Maps and Matrices

Heat maps offer an alternative pixel-based representation that encodes the data values by means of color. Fig. 2.5 shows a heat map of the same data depicted in Fig. 2.5, Fig. 2.2 and Fig. 2.3. The columns represent the homogeneous attributes (i.e. the classes). The rows represent the data items (i.e. the samples). Color represents the probability for each sample to belong to each class. The same color scale is used for all attributes, which is possible in case of homogeneous attributes that have the same nature and value ranges (probabilities in the depicted example). The rows and columns are ordered to reveal clusters of samples that exhibit similar class probabilities. The map can be augmented with *dendrograms* that explicitly depict hierarchical clustering of the rows (resp. columns), according to their similarities (Fig. 2.5).

Heat maps provide a clear overview of homogenous data. They help estimating the distribution of the values, and how it changes across different row and columns. As illustrated by Bertin [18], ordering has a vital impact on the patterns revealed by heat maps and matrices in general. Ordering simplifies the heat map by placing similar rows and columns close together, and reveals clusters of cells having similar values. Heat maps offer a good scalability with the number of attributes: About one hundred columns can be fit in one screen. On the other hand, heat maps offer fair scalability in the number of data points: few hundreds rows of can be fit in one screen, with minimum row visibility. Depicting larger number of items requires aggregation techniques to map multiple items into a 1-pixel row. Color is used by heat maps as the primary visual variable to show the values of homogeneous data. This limits the possibilities to use color for other purposes, such as showing selected items, or showing additional heterogeneous attributes of the data 1.2.

Instead of showing item-class relations (i.e. attribute values), heat maps and matrices can also be used to explicitly show relations between the attributes. In this case, both the rows and the columns of the map represent the attributes, with cell colors encoding their mutual relations.

Fig. 2.6 shows an example of a heat map depicting class confusions in the classification data of Fig. 1.1. The actual and predicted classes of the samples are mapped to the rows and columns respectively. The cells on the diagonal represent samples that are correctly classified in their actual classes. The remaining cells encode class confusions by color, with light color encoding fewer confusions. Other encodings of the class confusions are possible, such as cell size instead of cell color (Fig. B.2a).

Fig. 2.7 shows another example of a heat map depicting how multiple sets mutually overlap. The sets are mapped both to the rows and to the columns, with cell color encoding the overlap between the respective pair of sets.

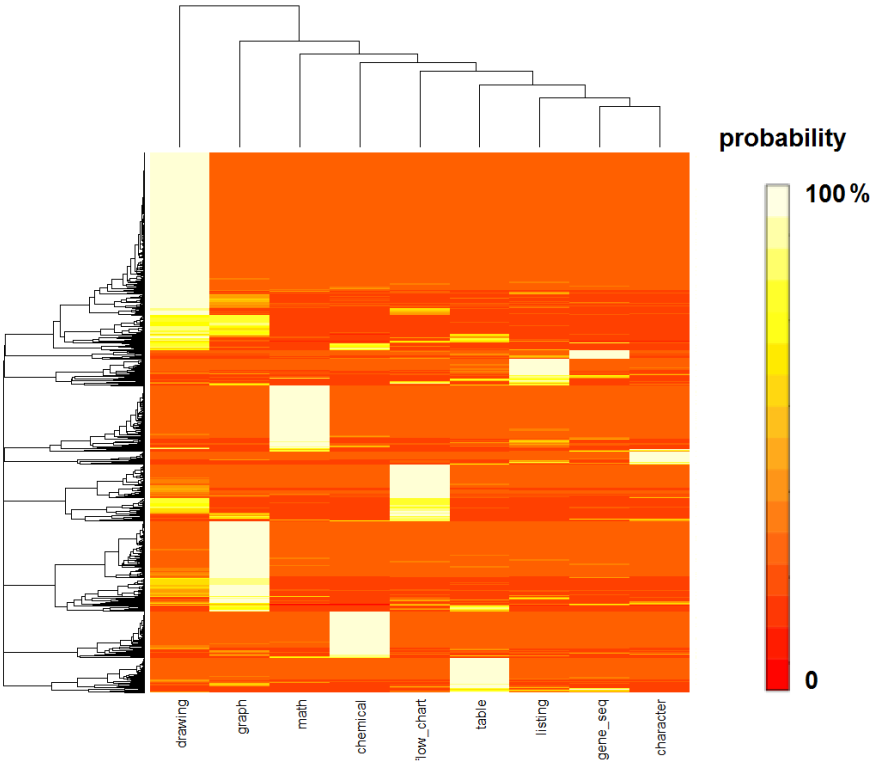


Figure 2.5: A heat map of the classification probabilities depicted in Fig. 1.1. Rows represent samples. Columns represent classes. Cell colors encode sample-class probabilities.

Actual vs. Predicted	drawing	chemical_structures	program_listing	gene_sequence	flow_chart	graph	math	table	character	FNs
drawing	244	18	2	1	49	12	3	3	0	88
chemical_structures	11	83	2	0	6	1	6	2	0	28
program_listing	1	0	12	1	6	5	0	1	0	14
gene_sequence_(dna)	0	0	6	7	0	1	0	10	0	17
flow_chart	13	0	0	0	87	3	0	2	0	18
graph	57	4	5	1	35	76	9	8	0	119
math	1	11	4	4	0	2	99	1	4	27
table	9	0	1	1	6	8	0	39	0	25
character_(symbol)	0	0	0	0	0	0	1	0	16	1
FPs	92	33	20	8	102	32	19	27	4	

Figure 2.6: A confusion matrix of the classification results in Fig. 1.1.

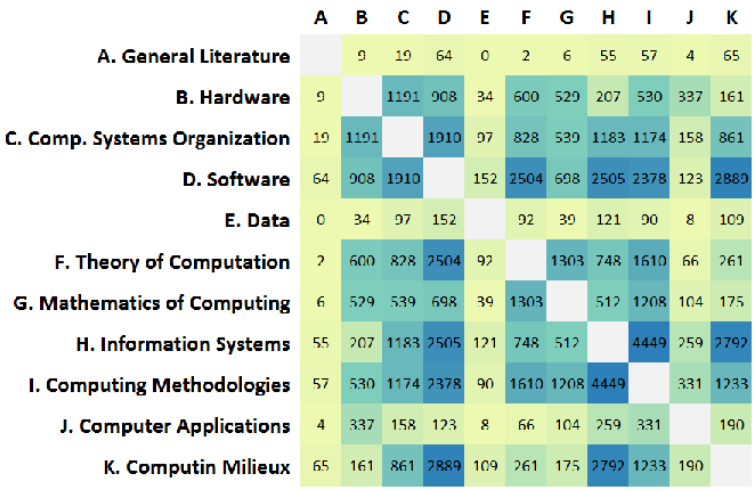


Figure 2.7: A heatmap showing how the 11 ACM classes [1] overlap as sets over ACM papers.

The matrix representation of pair-wise relations between the attributes gives a clear overview of value distributions by emphasizing these relations as primary objects in the visualizations. However, the information related to one attribute is rather scattered in multiple row and column cells. Besides, color is used as primary visual variable to show the attribute relations, which hinders the possibility of using it to show more details about their values or about other information about the data.

Node-link diagrams are alternative representations to matrix. They have been used to visualize multidimensional data, as explained next.

2.1.6 Ploceus: Network-based Visual Analysis of Tabular Data

Node-link diagrams can visualize relations between data values in a multidimensional dataset. The *Ploceus* [93] visualization system provides possibilities to create and interact with such diagrams. Fig. 2.8 shows the relation between two categorical attributes in a dataset: organizations (in orange) and funding programs (in blue).

It is also possible to contract the network into a one-mode network, which visualizes the relations between the values of one categorical variable. These relations can be analyzed in context of other attributes, by using small multiple. Fig. 2.8 shows collaboration between organizations on NSF grants, broken down by year and amount [93].

Node-link diagrams suffer from inherent limitation scalability. Depending on the network complexity (measured by number of edges and nodes), tens of nodes can be shown in one diagram. As the number of nodes and edges increases, the clutter caused by line crossings increases and hinders gaining insights in the data. Instead of using node-link diagrams to show single data items, these diagrams can be used to show aggregated similarity relation or correlations between the attributes, as explained next.

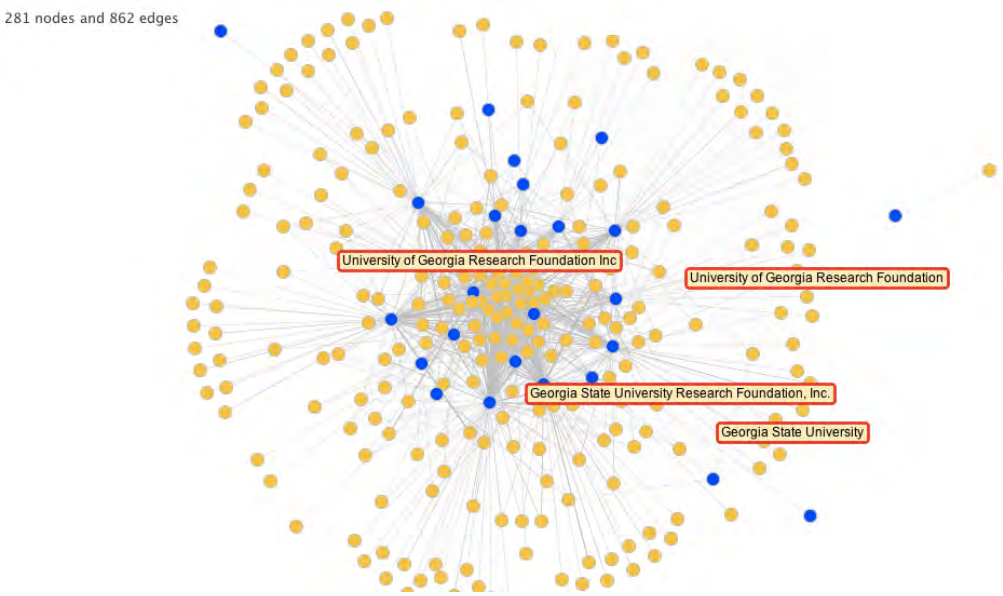


Figure 2.8: Ploceus view showing relation between the values of two attributes as a two-mode network [93].

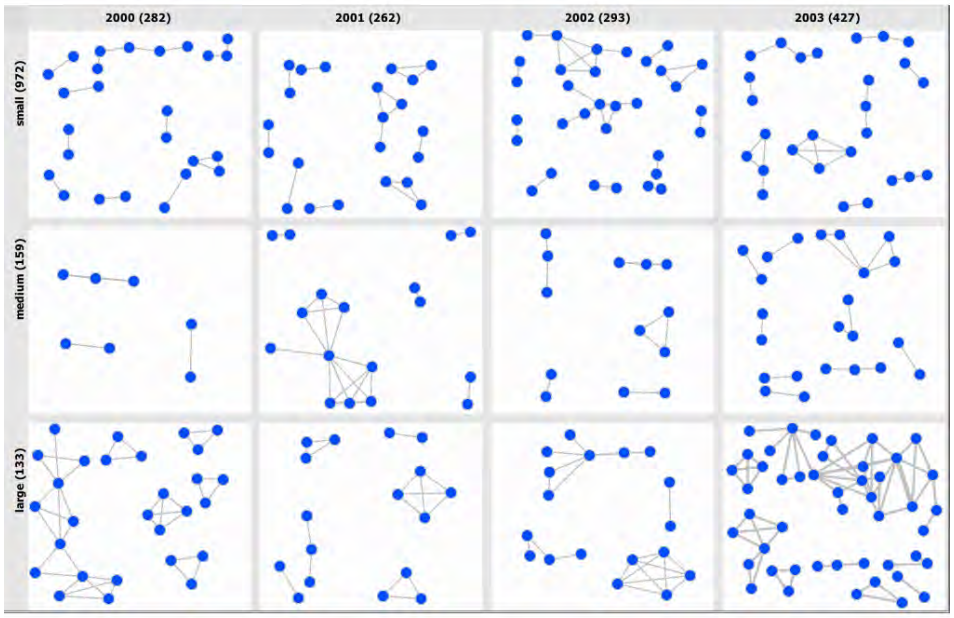


Figure 2.9: *Ploceus* view showing relation between values of one dimension [93].

2.1.7 Circos

The *Circos* system [87] is developed for analyzing genomic data by facilitating the comparison of genomes. It allows identifying similarities and differences between genomes by employing a variant of chord diagrams (Fig. 2.10). To analyze these similarities and to correlate them with the data properties, additional information is embedded in the radial layout using line charts and bar charts.

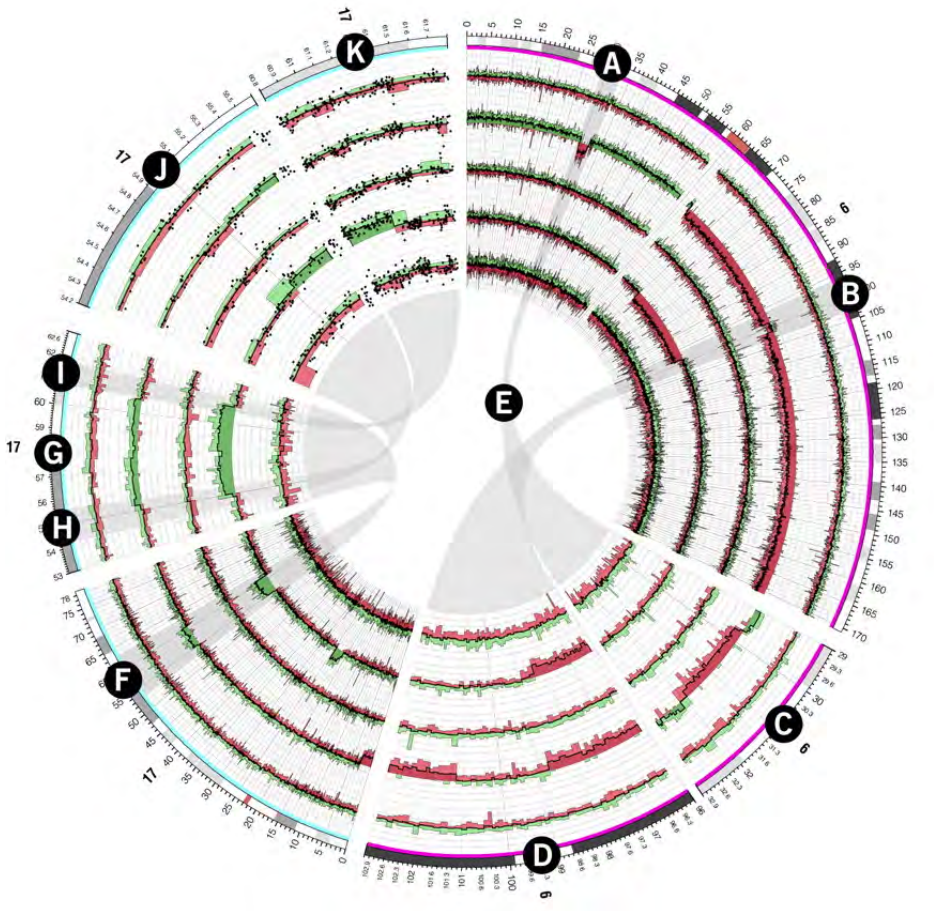


Figure 2.10: The *Circos* system [87] for visualizing genomic data.

Circos uses both geometric as well as graph-based techniques to visualize multidimensional data. It provides a variety of interactions to select certain genomes and to adjust the layout. While *Circos* is originally designed to visualize genomic data, it has been applied in other domain as well. In particular, *Circos* was demonstrated to be suited for visualizing relations between categorical variables, whentwo the number of categories is relatively smalls (tens of categories).

2.1.8 Summary

The previous sections give an overview on certain geometric, pixel-oriented, and graph-based techniques for visualizing multidimensional data. Icon-based techniques were not included in this overview due to their inability to handle large volumes of data, which is a primary goal of this thesis. Hierarchical techniques are also unsuited for homogeneous data, as they assign the attributes different degrees of visual primacy. On the other hand, visualizing homogeneous data requires techniques that treat the dimensions equally, as they are all of the same nature and generally of the same importance and relevance for the analysis.

Geometric techniques often suffer from clutter when they combine multiple dimensions in one plot or when they represent individual data items without aggregation. This is because projection can lead to overplotting or line crossings, depending on how the data points are represented. On the other hand, geometric techniques are potentially suited to show quantities of aggregated data values, e.g. using bar charts. This is because they use position and size which are effective visual channels for showing quantitative information.

Pixel-oriented techniques are efficient at providing overview of how the values are distributed in a dataset having a large number of items. Color can be used to show additional information about items that are laid out using geometric techniques. When used without aggregation, pixel-oriented techniques offer moderate scalability, depending on the number of available pixels. The insights gained in such visualizations depend heavily on how the items are ordered in the screen (which also applies to some geometric techniques as well).

Graph-based techniques are suited to show aggregated relations between a few number of dimensions or attributes, as in *Circos*. They can hence be used in combination with other techniques that show information about the data items.

As Chapter 3 demonstrates, the visualization technique for homogeneous data proposed in this thesis combines features from geometric, graph-based, and pixel-oriented techniques. It uses bar charts to show aggregated information about the data, visual links to show relations between the multiple attributes, and color to show further information about the items aggregated in the bar charts. Furthermore, the presented approach employs a multi-level overview+detail exploration environment to show different different aspect about the data at multiple levels of detail. Such use of multiple views has been proposed in various interactive visualization systems, as discussed in the next section.

2.2 Data Reduction

Visualizing large volume of multidimensional data comprising hundreds of thousands of items is challenging and quickly leads to issues with clutter if all the data values are presented at once in one screen. Therefore, it is important to reduce the amount of data to volumes that can be visualized effectively at a sufficient resolution and clutter. Keim [80] classifies data reduction methods into the following categories:

- **Dimension Reduction**: These techniques reduce the dimensionality by projecting the data on a low-dimensional space. Examples for that are Principal Component Analysis (PCA) [75] and *Multidimensional Scaling* (MDS) [86]. However, as illustrated in Sect. 2.1, these techniques are not suited for homogeneous data, as the results are hard to interpret in the projection space. To avoid impacting interpretability, data reduction should rather preserve the original homogeneous dimensions.

- **Sampling**: These techniques reduces the number of the data points by selecting a representative subset thereof. In fact, the data points available in a given dataset are usually a subset of all points that occur in a studied phenomenon. Sampling can be basically applied to reduce the volume of homogeneous data, especially if the number of data points is very large (e.g., in the order of millions) and would impose performance limits. However, when this is not the case, sampling should be avoided as it might mask important patterns such as outliers, and might cause certain queries and assumptions about the data to fail if the related data points are not included.

- **Querying (Filtering)**: Instead of randomly selecting a subset of the dataset, this subset can be described in a deterministic way. For example, a sales analyst might restrict the analysis to data from a certain region or shop. This avoids the issues of failed queries when using sampling, as the analyst is aware of what is included in the active subset. However, filtering limits the generalizability of insights and findings to the respective subset, as these might not necessarily apply to other subsets.

- **Segmentation**: As a divide-and-conquer strategy, segmentation allows dividing a large dataset into subsets that can be analyzed individually. As with filtering, this is usually done based on certain attributes, e.g., to divide sales data by region. This also means that the insights found are limited to the respective subsets of data.

- **Aggregation**: To reduce the volume of data involved in the analysis, it is possible to perform the analysis with groups of data points instead of individual points. This is performed by aggregating multiple data points in a group, based on similar characteristics. The visualization is restricted to visualize the resulting groups as visual elements of proportional sizes. Hence, the number of data points aggregated in each group can be infer from the visualization. It is also possible to retrieve these points on demand using interaction. The inclusion of all data points, even in an aggregated form, is advantageous: Points that match a specific user query are always retrieved and highlighted in the visualization, which does not always hold with the other data reduction methods.

Other methods for data reduction include statistical methods that summarize a large amount of data by appropriate model parameters. The approach adopted in this thesis is based mainly on data aggregation and partially on filtering, as explained in Chapter. 3.

2.3 Coordinated and Multiple Views

One way to cope with the complexity of visualizing multidimensional data is to display the data in multiple views using different visualizations. These views show different aspects about the same data, and can be linked together using *brushing and linking* techniques [30, 122]. Several systems use such *coordinated and multiple views (CMV)* to enable rich visual analysis of multidimensional data in different domains such scientific simulation [38], finance [25], and GeoAnalytics [72]. Also, several models and architectures were devised for building CMV system [19, 32, 108, 109]. Roberts [122] surveyed different models and techniques for implementing such views and the challenges of applying them in *VA*. He discusses many aspects of implementing CMV system, including data processing and oreparation, view generation and multiple views, exploration techniques, coordination and control, tools and infrastructure, human interface, and usability and perception.

Fig. 2.11 shows an example of a CMV system called *ComVis*. It integrates various views that show different aspects of the data, including bar charts, histograms, line charts, heat maps, scatterplots, parallel coordinate plots, and sortable tables. Brushing items in one view updates the selected items in the other views, highlighted in red.

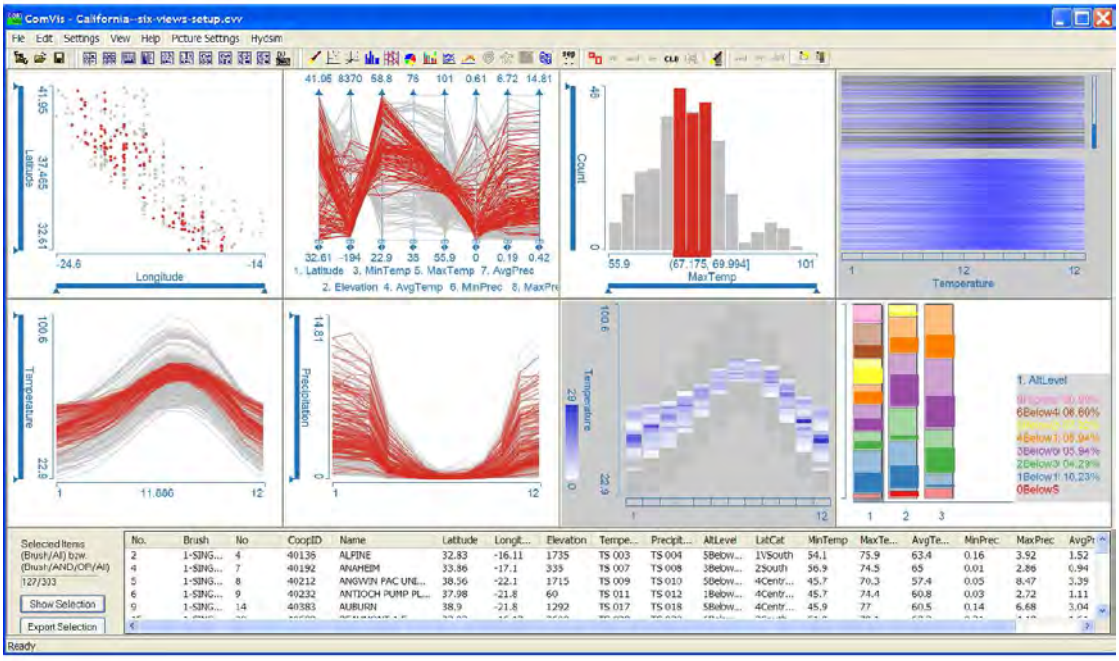


Figure 2.11: Coordinated and multiple views in *ComVis* [97].

Fig. 2.12 shows another example of a CMV system called *Improvise*. This system is designed for building highly-coordinated visualizations, by means of a shared-object coordination mechanism coupled with an expression-based visual abstraction language.

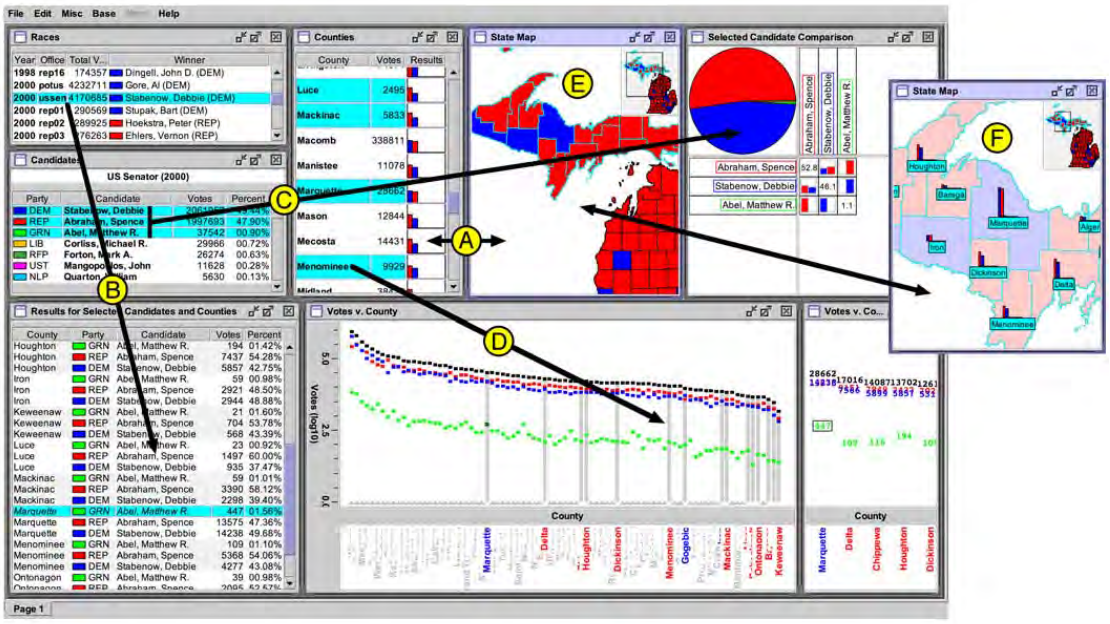


Figure 2.12: The *Improvise* [164] for building highly-coordinated visualizations.

The use of multiple views enables an *overview+detail* paradigm for exploring multidimensional data. This paradigm follows Shneiderman's *visual information-seeking mantra*: "overview first, zoom and filter, then details-on-demand" [139, p. 2]. It uses *spatial separation* as the main scheme to allow users to work at, and move between, focused and contextual views [30].

An overview+detail exploration environment is well-suited to support visual analysis using aggregation-based visualizations. It allows to interactively explore certain aggregated elements at multiple levels of detail. Each level of detail has its specific view which is linked with the other views to update the information it shows upon changes in the selection. Appropriate brushing and linking techniques are need to select elements of interest, e.g., using mouse interaction, and to highlight them in linked views, e.g., using color.

The Visual-Analytic approach proposed in this thesis follows the above-mentioned paradigm to analyze homogenous data. It uses both aggregation and overview+detail exploration, as explained in detail in the next chapter.

CHAPTER 3

The Proposed Approach

This chapter outlines the approach adopted in this thesis for analyzing homogeneous data [1]. After motivating the use of a *VA* paradigm, I explain how *VA* is applied to computationally analyze and aggregate the data (Sect. 3.2). Then, I propose a novel visualization that is suited to display and interact with the aggregated data (Sect. 3.3). Sect. 3.4 elaborates on the design of the interactive environment for exploring the data at multiple levels of detail.

I close this chapter by comparing the proposed approach with the state-of-the-art methods presented in Chapter. 2, and emphasizing its scalability advantages and limitations (Sect. 3.6). I also report an overview of qualitative and quantitative evaluation results of the proposed design, applied to the different data classes (Sect. 3.7).

3.1 Visual-Analytics Paradigm

Classical visualization methods suffer from inherent limitation: they cannot always depict more data than pixels available on the screen. Depending on the visual representation used, and the complexity of the data depicted, the visualization might become quickly cluttered even with hundreds of data items. The field of Visual Analytics [151] aims to intertwine automated analysis techniques with interactive visualizations enriched by cognition and perception to allow solving tasks and gaining insights that are cumbersome to perform using either paradigm independently [79]. Automated methods exploit the computational capabilities of modern computers to find the desired information and patterns in huge amounts of data. Interactive visualizations allow involving the user in the analysis loop and incorporating domain knowledge to guide the analysis process and toassess the results. Moreover, the human visual system is excellent at detecting unexpected patterns and anomalies that are hard to be captured in advance by automated methods.

[1] Part II elaborates in detail on the specific analytical methods used to aggregate the data for each of the three class of homogeneous data addressed in this thesis. It also explains how the generic visual metaphor and exploration environment are customized to visualize the data and solve related tasks for each of these classes.

The *VA* paradigm proposed in this thesis follows Keim's *VA* Mantra [79, p. 82]: "*Analyse first – show the important – zoom, filter and analyse further – details on demand*". The next three sections show how analytical methods compute the important information in homogeneous data, how the visualization is designed to show this information, and how the interactive exploration environment allows conducting further analysis and obtaining details on demand.

3.2 Automated Analytical Methods

Automated analysis helps in analyzing large amount of data by computing the important information. This reduces the volume of data and makes it appropriate for visualization and exploration. The next sections give a unified overview of how automated analysis is applied to three classes of homogeneous data. These sections cover five common aspects of data analysis that arise in all of the three classes of data: (1) associations within the data (2) aggregation (3) filtering (4) relations between the homogeneous attributes (5) correlation with other attributes. Further details about the analysis of each of these classes are provided in Part II.

3.2.1 Entity Association with Homogeneous Attributes

A major analysis task of homogeneous data is to understand how the row entities differ in their relation with the homogeneous attributes. These relations can be expressed by means of an association function $r(e_i, A_j)$ between the entities $e_{1\leq i\leq n} \in E$ and the attributes $A_{1\leq j\leq m}$. This function differentiates between data points that have high or low association with an attribute A_j, depending on the class of homogeneous data being analyzed:

- **Probabilistic classification data**: The association values with attribute A_j are equal to the probabilities computed for the samples in the respective class:

$$r(e_i, A_j) = A_j(e_i) \tag{3.1}$$

  The association with attribute A_j is hence larger for samples that are more likely to fall in the respective class.

- **Element-set memberships**: The association value of an element e_i with attribute A_j is based on the *set membership degree* of e_i [8], and whether e_i is an element of the respective set S_j represented by attribute A_j:

$$r(e_i, A_j) = \begin{cases} (m - degree(e_i))/(m-1) & e_i \in S_j \\ -1 & otherwise \end{cases} \tag{3.2}$$

  where $degree(e) = |\{S_j : 1 \leq j \leq m \wedge e \in S_j\}|$ indicates the number of sets to which this element belong. Elements e_i that are exclusive to a set S_j have a degree of 1, and hence are assigned the highest association value $r(e_i, A_j) = 1$ with attribute A_j. On the other hand, elements that exist in all the sets have a set membership degree of m

and are hence assigned the lowest association value of 0 with all attributes, as they are not specifically associated with any of these attributes. Elements that are not members of set S_j are assigned a negative association value of -1 with attribute A_j, and can hence be excluded when analyzing the elements of this set.

- **Categorical data**: The association value of an element e_i with attribute A_j is based on comparing the actual frequency $A_j(e_i)$ of the category combination (e_i, A_j), and the expected value e_{ij} of this frequency under $H0$, assuming 0 association (Eq. C.1):

$$r(e_i, A_j) = tanh\left(\frac{A_j(e_i) - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij} \cdot \left(1 - \frac{f_{i+}}{n_{++}}\right) \cdot \left(1 - \frac{n_{+j}}{n_{++}}\right)}}\right) \tag{3.3}$$

  where $n_{i+} = \sum_{j=1}^{m} A_j(e_i)$ and $n_{+j} = \sum_{i=1}^{n} A_j(e_i)$ are the marginal row- and column frequencies in the homogeneous data table, and n_{++} is the sum of all table frequencies.

  Sect. C.2.1 provides more details about the associations measures used in Eq. 3.3. The association is positive when the frequency is higher than expected under $H0$, with +1 being the largest positive association value. Likewise, the association is negative when the frequency is lower than expected under $H0$, with -1 being the largest negative association value (e.g., when the respective row and column categories never appear together).

3.2.2 Aggregating Homogeneous Data

The proposed approach aggregates the values of each homogeneous attribute $A_{1 \leq j \leq m}$ individually. The association values $r(e_i, A_j)$ are used to divide the entities into b groups $G_{j1}, ..., G_{jb}$ with respect to attribute A_j. This is performed by binning the value range of r into b bins $B_1 = [low_1, high_1]$ .. $B_b =]low_b, high_b]$, as illustrated in Fig. 3.1. Entities that fall in the same bin $B_{1 \leq k \leq b}$ are aggregated together in one group G_{jk}:

$$G_{jk} = \{e \in E : r(e, A_j) \in B_k \subset \mathbb{R}\} \tag{3.4}$$

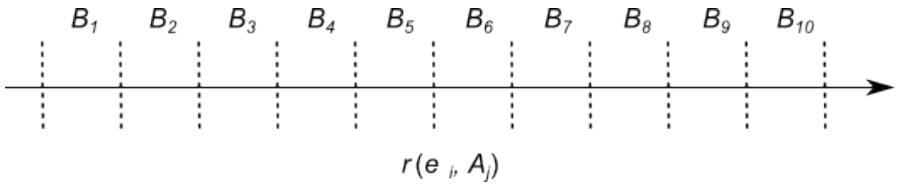


Figure 3.1: Binning the value range of the association function with attribute A_j.

The choice of the number of bins b and the value ranges of the bins B_1 ... B_b depend on the class of homogeneous data being analyzed, on the data distribution, and on the tasks to be solved based on the data:

- In case of **element-set memberships**, it is possible to define the bins to be the set-membership degrees, so that elements of the same degree in a set fall in the same bin. It is also possible to define bins of different ranges, for example, to group the elements into those that are exclusive to their sets (degree(e) = 1), those that are shared between 2-4 sets ($2 \leq degree(e) \leq 4$), 5-9 sets ($5 \leq degree(e) \leq 9$), and between 10+ sets $degree(e) \geq 10$.

- In case of **probabilistic classification data**, equal bin sizes should be used, which allows depicting the bins over a linear and equally-divided probability axis. It is possible to restrict the bins to the effectively-used probability range $\subseteq [0, 1]$ instead of spanning the whole range $[0, 1]$. It is also possible to aggregate samples classified with 100% probability in a special bin, to aid easy access to these samples. The number of bins can be selected to be equal to the number of classes m. It can also be selected according to the probabilistic classification algorithm used, e.g., a multiple of k in case of k-NN classification.

- In case of **categorical data**, a fixed number of equally-sized bins is used to show the proportion of entities at multiple levels of associations with each column category. This number can be determined by statistical rules such as Scott's normal reference rule [131].

Aggregating homogeneous data in $m \times b$ bins enables a massive reduction of the data volume, compared with the number of row entities n. Each entity is aggregated in up to m different bins, according to the m attributes individually. As explained in Sect. 3.4, a list of the entities aggregated in a specific bin can be obtained interactively for further analysis.

It is also possible to aggregate homogeneous data "vertically" along the homogeneous attributes, instead of the row entities. For example, when analyzing set-typed data, it is possible to aggregate multiple sets into one set, using set union. This allows focusing the analysis on certain sets without ignoring the remaining sets that can be aggregated in one group, e.g., "others". Similar aggregations can be applied to frequency and probabilistic data by summing up frequencies or probabilities cell-wise to combine the respective column categories or classes.

3.2.3 Filtering Homogeneous Data

Besides aggregating the entities, it is also possible to filter out certain entities either to hide them from the visualization or by completely excluding them from the analysis. This enables reducing the amount of information displayed and exploring the remaining entities at greater detail. The entities e can be filtered out based on their associations $r(e, A_j)$ with certain homogeneous attributes A_j or on other criteria. Each of the classes of homogeneous data can employ filtering for specific purposes:

- When analyzing **element-set memberships**, it is possible to filter elements by their set-membership degrees. For example, filtering out elements that ar eexclusive to their sets allows focusing on elements involved in set overlaps. This is especially useful when the sets in a set system do not exhibit a lot of overlap, leaving the majority of elements belonging to one set only. On the other hand, when a set system exhibits a high degree of

overlap, it is possible to filter out elements that belong to the majority of sets. This enables focusing on elements that are specific to certain sets or set combinations and analyzing if they have specific features compared with other elements.

- When analyzing **probabilistic classification data**, it is advantageous to filter out non-class samples that are classified as such, especially when they are assigned low probability in this class. Such true negatives usually form the majority of the samples, and are less relevant when analyzing classification results. Filtering out these samples enables focusing the analysis on misclassified samples that are more relevant for improving the classification performance.

- In case of **categorical data**, it is possible to filter out row entities that have low overall support in the database (i.e. low marginal row frequency). The associations computed for these entities are usually less significant (e.g., based only on one occurrence in the database). Filtering out these entities enables focusing the analysis on significant associations that imply generalizable association rules.

Instead of filtering data points, it is also possible to filter certain homogeneous attributes. This allows analyzing associations between a subset of these attributes, and how they relate to the data points. This is useful when the number of attributes exceeds the visual limits of the visualization design. For example, when analyzing set-typed data, it is possible to restrict the analysis to a small number of sets to find out how they overlap, regardless to the remaining sets. These relations between the attributes is a central information in the data, as explained next.

3.2.4 Relations Between Homogeneous Attributes

The homogeneity of the data described by homogeneous attributes enable new forms of computing the relation between the real-world entities they represent, beyond the general case of multidimensional data. In the general case, the relation is computed based on statistical measures such as the covariance between the attributes by treating them as dimensions. Alternatively, these relations can be computed to address the specific nature and characteristics of the data, and how the comparison tasks are defined. This facilitates the interpretation of these relations and what the computed measures represent. The proposed approach employs different measures to compute the relation between attributes, depending on the class of homogeneous data they represent (the relation is denoted as $rc(j_1, j_2)$):

- **Set similarities:** When the homogeneous attributes represent sets, their pair-wise relations can be computed as similarities between these sets. Many set-related tasks involve finding pairs of sets S_1 and S_2 that exhibit higher similarity than other pairs, with respect to the number of elements shared between them $|S_1 \cap S_2|$. Several similarity measures between finite sets have been proposed in the literature [8]. The choice of an appropriate measure depends on the data characteristics and on the tasks to be solved. Jaccard [55] proposed a symmetric similarity measure:

$$Jaccard(S_1, S_2) = |S_1 \cap S_2| / |S_1 \cup S_2| \tag{3.5}$$

Though this measure is intuitive to interpret, it is, however, biased with respect to the set sizes: Larger sets have higher chance to overlap, and hence are systematically assigned high pair-wise similarities. One way to address this issue is to compute the similarity based on overlap probability of the sets. This can be performed by treating the sets as events and computing the deviation between the actual probability and the probability in case of conditional independence. A normalization using the χ^2 statistic can be used to eliminate bias toward set sizes (Eq. 3.6):

$$rc(j_1, j_2) = \frac{P(S_{j_1} \cap S_{j_2}) - P(S_{j_1}) \cdot P(S_{j_2})}{\sqrt{P(S_{j_1}) \cdot P(S_{j_2})}} \tag{3.6}$$

Other measures of set similarity can be employed such as Tversky [155] generalized index:

$$Tversky(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cap S_2| + \alpha \cdot |S_1 \setminus S_2| + \beta \cdot |S_2 \setminus S_1|} \tag{3.7}$$

It is also possible to weight shared elements differently when computing the similarity, instead of just using their count. For example, elements of degree 2 in $S_1 \cap S_2$ can be weighted higher than other elements, as they belong exclusively to the overlap. Also, elements exclusive to their sets can be excluded from the computation of the denominator. This allows focusing on shared elements especially in sparse family of sets that exhibit little overlaps.

It is also possible to compute the similarity between multiple sets $S_1, S_2, \cdots, S_k$ based on their intersection $S_1 \cap S_2 \cap \cdots \cap S_k$. The above-mentioned measures can be generalized to accommodate multiple sets, e.g., using $|S_1 \cup S_2 \cup \cdots \cup S_k|$ for normalization and $P(S_1) \cdot P(S_2) \cdot ... \cdot P(S_k)$ for computing the probability in case of conditional independence.

- **Class confusions**

  In probabilistic classification, the classifier returns the class with the highest probability as the predicted label $l^p(e)$ for a given input sample $e \in E$ [7]:

  $$l^p(e \in E) = \underset{1 \leq j \leq m}{\operatorname{argmax}} A_j(e) \tag{3.8}$$

  The confusion between classes j_1 and j_2 is computed by comparing the predicted labels $l^p(e)$ with their actual labels $l^a(e)$:

  $$rc(j_1, j_2) = |e \in E : (l^a(e) = j_1) \wedge (l^p(e) = j_2)| \tag{3.9}$$

  The relation between two attributes representing two classes is based on the confusion between these classes. This relation is generally asymmetric because when samples of one class are confused with another class, the opposite does not necessarily happen [7].

- **Categorical similarities**

  When the homogeneous data represent a contingency table, the similarity between two columns is computed based on how similar their relations with the row entities are. Two

columns are high relation if the row entities exhibit positively (or negatively) high association with both of them (Eq. 3.3). One way to express this relation is to compute the correlation between these row-column association values:

$$rc(j_1, j_2) = \frac{1}{n_{+j_1} + n_{+n_2}} \cdot \sum_{e \in E} r(e, A_{j_1}) \cdot r(e, A_{j_2}) \tag{3.10}$$

Correspondence Analysis offers another possibility to compute the relations between the attributes [52, 126]. It is designed to compute a 2D reduction of contingency tables in a similar way as with *Principal Component Analysis*, and can compute the similarity between the table dimensions according to their cell values.

3.2.5 Correlations with non-Homogeneous Attributes

As explained in Sect. 1.1, the datasets containing homogeneous attributes usually exist along with other heterogeneous attributes. The methods proposed in this thesis treat homogeneous data as the primary focus of analysis. Nevertheless, many analysis tasks involve analyzing the heterogeneous attributes in relation to the homogeneous data.

One generic analysis task is to compare the attribute distributions between different groups of data. For example, users often want to know if elements that belong to a specific set or sets combination have different values of their other attributes than the remaining elements. The same applies when analyzing classification data, to figure out if high probability to belong to class is associated with certain attributes values.

One way to support the comparison of numeric attribute values between two groups of entities $X1 \subseteq E$ and $X2 \subseteq E$, is to use a significance statistic. Welch's t-statistic [34] is suited when the groups are of unequal sizes. The statistic for each feature f_k can be computed as follows:

$$t_k = \frac{mean(f_k(X_1)) - mean(f_k(X_2))}{\sqrt{var^2(f_k(X_1))/|X_1| + var^2(f_k(X_2))/|X_2|}} \tag{3.11}$$

The corresponding *p*-values according to Student's *t*-distribution should be used to enable a direct comparison of the resulting *t*-values. Features with more significant differences between X_1 and X_2 will have lower *p*-values. It is also possible to infer whether the differences are significant or not by applying a *t*-test to the values, to figure out if the differences can be explained by random sampling only.

Binning is another way to compare attribute distributions. This is not only suited for a numerical attribute, but also for a categorical attribute as categories naturally define bins. The χ^2 statistic is suited to compare two binned histograms having b bins:

$$\chi^2_k = \sum_{l=1}^{b} \frac{(f_k(X_1) - f_k(X_2))^2}{f_k(X_2))} \tag{3.12}$$

This is suited to compare a set of selected entities X_1 (i.e. observed behavior) against the remaining entities X_2 (expected behavior), as used for testing the *goodness of fit* between observed and theoretical distributions. In contrast with the *t*-statistic which is a holistic mean and variance,

binning allows capturing differences that are masked by holistic measures. For this purpose, an appropriate number of bins should be used, based on the underlying feature distributions. Other statistics that are independent of the distributions can be used to compare two numerical distributions such as the Kolmogorov-Smirnov statistic [96].

Another task related to the additional attributes is to find out if two or more groups of entities can be separated (i.e. distinguished from each other) by means of an attribute. This task arises, for example, when analyzing classification data in order to find if certain confusions between true positive and false positives can be resolved by means of data features extracted from the samples. For this purpose, separation measures can be used to assess the separation power of individual attributes or attributes combinations. Variants of the F-Measure [115] can be used for this purpose, applied to histograms h_{1k} and h_{2k} of a feature f_k in X_1 and X_2 respectively (where b is the number of histogram bins):

$$F_k = \sum_{b=1}^{b_h} h_{1k}(b) \cdot \frac{h_{1k}(b)}{h_{1k}(b) + h_{2k}(b)} \tag{3.13}$$

The above-mentioned tasks can be generalized to analyze the behavior of pairs of attributes among different subsets of the data. Other attribute-related tasks can also be defined for specific instances and applications. It is important to choose appropriate analytical techniques that fit (1) the task, e.g., finding differences vs. separability, and (2) the data characteristics, e.g., numerical vs. categorical attributes, specific attribute distributions, etc. For these purposes, many other data analysis techniques can be applied beyond the above mentioned exemplary ones.

3.2.6 Summary

The previous sections demonstrate how automated analysis can be applied to compute associations within homogeneous data tables and relations between the homogeneous attributes, as well as correlations with other heterogeneous attributes in the dataset. The previous sections also demonstrate how automated analysis is vital to reduce large volumes of homogeneous data by means of aggregation and filtering.

It is essential to mention that the presented techniques exploit the homogeneity of the data which offers new analysis possibilities that are not possible to use in the general case of (heterogeneous) multidimensional data. This was demonstrated when computing the associations (e.g., based on set-membership degrees) and the attribute relations (e.g., set similarities), as well as when aggregating and filtering the data (e.g., based on set-membership degrees, or using set operations).

Part II gives further details about how automated analysis is applied for each class of homogeneous data addressed in this thesis. Employing a common analysis paradigm make the outcome of automated analysis similar for all these classes: binning-based aggregated associations, as well as quantified attribute similarities. This enables developing a common visual metaphor to visualize this outcome, as the next section illustrates.

3.3 Visual Representation

Despite the different nature of the data classes addressed in this thesis, they share several commonalities that allow developing a common visual metaphor for them. This is possible, thanks to the computational analysis which creates a unified abstract representation of the data. The association measures introduced in Sec. 3.2.1 convert the values of a homogeneous data table into associations in the range $[0, 1]$ or $[-1, 1]$, depending on the measure used. These associations $r(e_i, A_j)$ have similar interpretation: strong / weak association between the entities $e_i \in E$ and the respective column category A_j. This enables to visually treat these values in a similar manner, unlike the raw data that have different interpretations, i.e. binary set memberships, probabilities, and frequencies. Moreover, this enables aggregating these association values similarly into binned histograms (Sect. 3.2.2). The histogram bins differentiate between row entities that have strong or weak associations with the respective columns.

The wheel metaphor

This thesis proposes a common visual metaphor called the *wheel metaphor* to visualize homogeneous data. This metaphor consists of the following components (Fig. 3.2):

- The attributes $A_1, ..., A_m$ are represented as sectors of a ring chart, with the attribute labels being depicted next to these sectors.
- The aggregated associations $rc(e_i, A_j)$ are depicted as binned histograms into these sectors along the radial dimension. The outermost histogram bin represents entities having highest association with the respective attribute. The innermost histogram bin represents entities having the lowest association with the respective attribute. The length of histogram bar k in sector j represents the number of entities $|G_{jk}|$ at association level k with attribute A_j (Eq. 3.4).
- The relation between the attributes $rc(j_1, j_2)$ are represented as links between the sectors, depicted in the inner circle ring chart. The thickness of a link encodes the relation (e.g., similarity) between the respective attributes.

The ring sectors are reordered to optimize the visual layout by reducing clutter and grouping related attributes. This is done by selecting an order that makes thicker links as short as possible, which has two advantages:

- Reducing the ink used to depict the links and the clutter caused by crossings between thick links.
- Placing sectors of attributes having strong mutual relations close together, to reveal groups of inter-related attributes.

Computing an optimal ordering for the above criteria is an NP-Complete problem [95]. In this thesis, an $O(m^2)$ greedy algorithm is proposed to produce an approximate solution (where m is the number of attributes), as explained in detail in Sect. A.3.1.

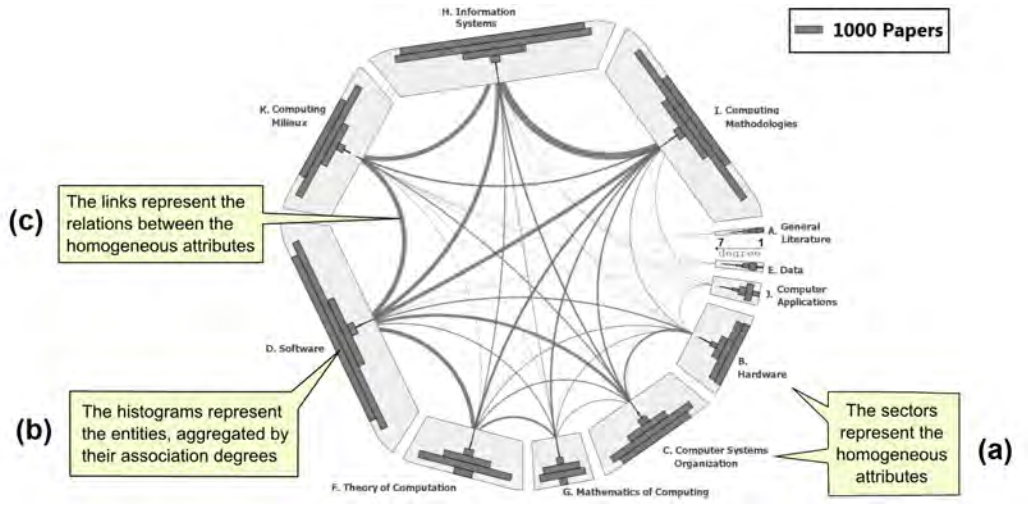


Figure 3.2: The proposed visual metaphor for visualizing homogeneous data: (a) sectors for attributes, (b) histograms for aggregated entities, (c) links for attribute relations.

The design rationale of the wheel metaphor is to depict the attributes as primary objects of the visualization (as labeled sectors), which contain all related information about these attributes. This information encompasses entity associations (depicted as histograms), and relations with other attributes (depicted as links). This is because the homogeneous attributes are central objects in the analysis and should be easily identifiable, along with all related information. There are two main reasons for choosing a radial layout for the attributes:

- The inner circle of the layout is a natural and compact place to depict the relations between the attributes as links. Placing the sectors in a grid would hinder depicting the arcs. Placing the sectors in one linear row imposes an artificial ordering on the attributes since the row has a start and an end. It also requires about π times more space in width (due to circle unfolding) and is unsuited to show the links as they would vary highly in length. Visualization expert, Stephen Few, comments on circular data visualizations [45]:

  "One of the few things that they can display well are the relationships between a single list of items ... The circular arrangement of a single list of comparable items ... lends itself nicely to the display of relationships between them, represented by lines that connect them. In this case, a linear arrangement would not work as well."

- The placement of the histograms simulates the magnet metaphor of Yi et al. [173], with sector labels acting as magnets on the radial dimension. The items aggregated in the outermost bars (closer to the labels) are highly associated with the respective attributes, and have a weak or no association with the other attributes. These items represent, for example, set-exclusive elements, or samples having high probability $\geq 90\%$ to belong to a class. The items aggregated in the innermost bars (closer to the inner circle) are less

associated with the respective attributes, as they have association with the other attributes. For example, a sample having 30% probability to belong to a class, has 70% probability to belong to the other classes, and is hence attracted to the middle area that represents shared elements. Likewise, an element in set S that exists in most other sets in a set system is aggregated in the inner bins. This element is less attracted by the magnet representing the label of S, as it is closer to the inner circle which attracts shared elements.

- The depiction of the histograms and the links in one view facilitates a better interpretation and richer analysis. Each element can be aggregated into multiple histograms in different sectors (e.g., being a member of multiple sets). Links explicitly reveal this redundancy by encoding the existence of shared elements that contribute to the relation between two homogeneous attributes (e.g., overlap relation between the sets). Additionally, integrating links and histograms in one view enables to visually relate these pieces of information. For example, when the link encodes a confusion between two classes, it is possible to infer which probability the confused samples are assigned in both classes. Interaction is important to perform this kind of analysis, as explained in Sect. 3.4.

The wheel metaphor combines both geometric and graph-based techniques to represent multidimensional data, according to Keim's classification [81] (Sect. 2.1). It uses position and size to represent aggregated entity associations with homogeneous attributes, as well as links to show relations between these attributes.

Use of color The metaphor as presented so far uses only spatial and size variables to represent the data. This allows using color to support brushing and highlighting of selected elements, as illustrated in Fig. 3.3. For a given selection the entities $E_{sel} \subseteq E$, it is possible to compute the fraction of selected elements in the elements G_{jk} aggregated in each histogram bar (Eq. 3.4). The selection can be indicated by coloring selected fractions of the histogram bars. Likewise, it is possible to compute the fractions of attribute relations contributed by the selected elements, e.g., portion of selected elements in an overlap between two sets. These fractions can also be highlighted in the links.

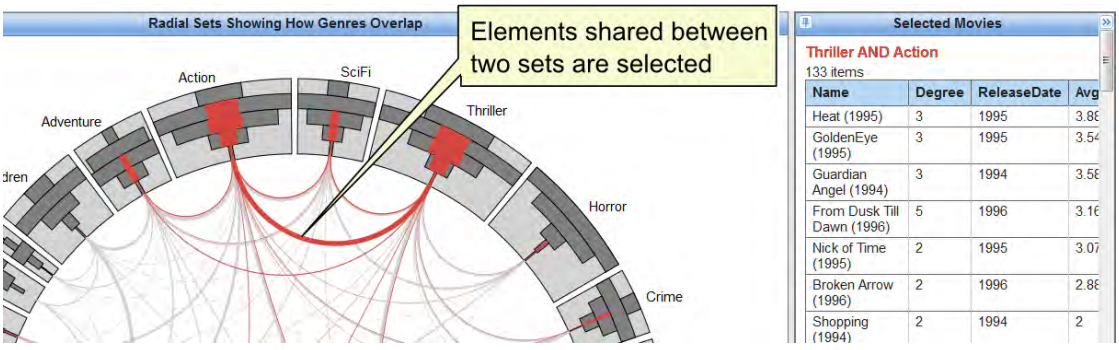


Figure 3.3: Color indicates fractions of bars and links representing selected elements.

Color can also show additional information about the data in the histograms, such as values of an attribute f_k for elements aggregated in the bars. In case the attribute f_k is categorical, each bar can be divided into colored sections to show the breakdown of its elements by the respective categories (Fig. 3.4a). In case the attribute f_k is numerical, the bar color can be used to encode an aggregation of the attribute values in the bar elements, such as the average or the median of these values (Fig. 3.4b).

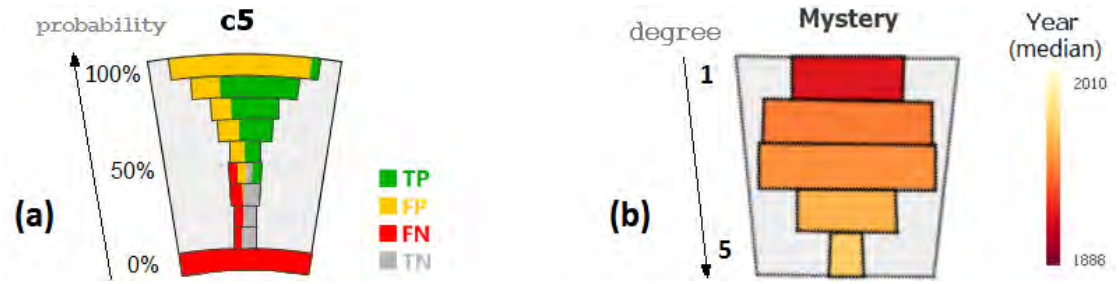


Figure 3.4: Coloring histogram bars by (a) a categorical attribute indicating classification results (Chapter. B), (b) a numerical attribute (only the medians per histogram bar is shown).

The attribute values can alternatively be explored in additional views using suited visualizations, such as bar charts and boxplots (Fig. 3.5). These views can also show selected fractions of the elements. Interaction allows exploring the elements aggregated in certain bars in detail, along with their attributes as explained in the next section.

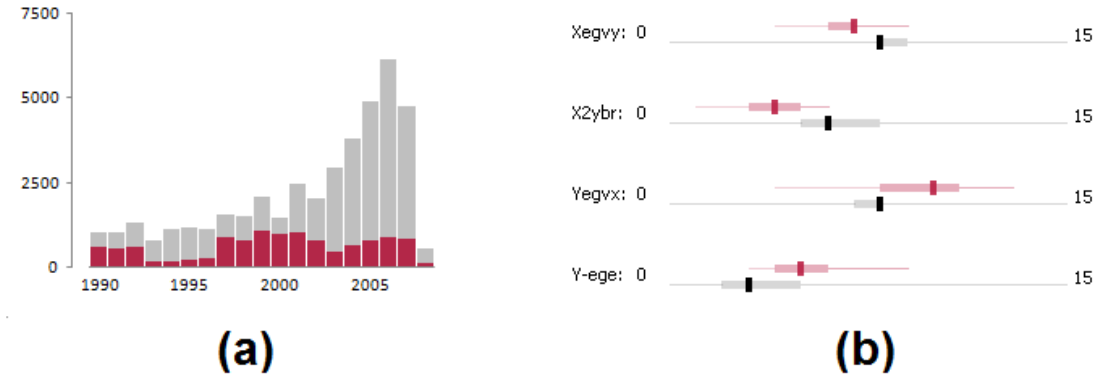


Figure 3.5: Visualizing attributes in additional views (a) a bar chart of an attribute with fractions representing selected entities highlighted in color. (b) box plots to give an overview of multiple attribute distributions and compare them among selected and unselected entities.

3.4 Interactive Exploration Environment

The wheel metaphor is designed to show a rich overview of homogeneous data, which provides information both on entity-attribute associations and on attribute relations. For this purpose, the metaphor makes use of aggregation by showing groups of entities (Eq. 3.4) instead of individual items. To enable exploring individual items aggregated in a certain group, multiple-view exploration is needed (Sect. 2.3).

The structural similarity between the classes of homogeneous data, and the unified abstraction employed by the wheel metaphor, both allow (re-)using similar exploration environments for all of these classes. Depending on the concrete tasks associated with each class, the respective environment might still contain specific views that support these tasks.

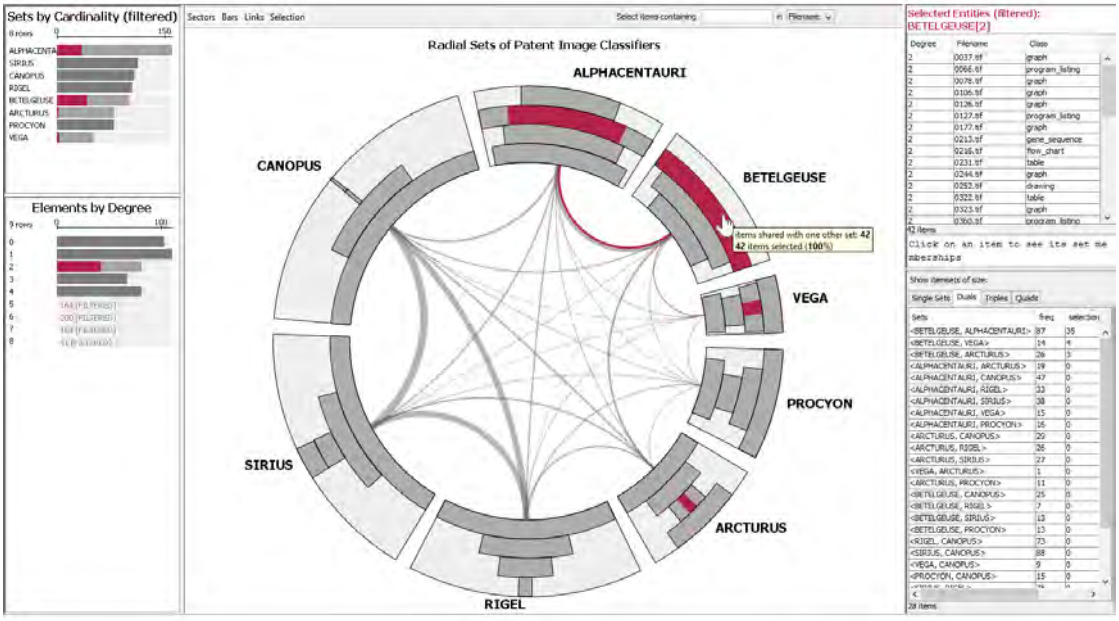


Figure 3.6: The interactive exploration environment with customized views for analyzing set overlaps. The two views on the left-hand side provide a summarized overview of the dimensions (top-left) and the entities (bottom-left). The central view shows gives a richer overview of the dimensions and the entities at once using the wheel metaphor. The right-hand side view shows detailed information about selected entities as in tabular lists.

The basic exploration environment, shared between all classes of homogeneous data, is designed following Shneiderman's *visual information-seeking mantra*: "overview first, zoom and filter, then details-on-demand" [139]. Fig. 3.6 shows an example exploration environment for exploring overlapping sets. The left-hand side views show a high-level overview of the data, showing the sets by their cardinality and elements by their degrees. The central view shows a more detailed overview of the data using the wheel metaphor. The right-hand side views show details about selected elements on demand, including a list of these elements and their attributes,

as well as fractions they make up of the set overlaps. Part II explains in detail how the exploration environment is customized for three classes of homogeneous data, in particular which specific views and interactions are included to aid the analysis of the data.

Interaction is vital to explore the data at multiple levels of detail. It allows selecting certain elements in one view, and checking these elements in the other views. The elements can be selected by clicking on the respective visual items. In Fig. 3.3, the elements shared between two sets are selected by clicking on the respective link. In Fig. 3.6, the elements of degree 2 in a set are selected. In both cases, all occurrences of these elements in the wheel view and in the other views are highlighted. A textual description of the query used to select the elements is displayed in red at the top of the detail views. The selected elements can be refined further by iteratively refining this query. This can be done by combining multiple selections using set operations (set union, set intersection, and set difference). For example, it is possible to exclude elements in set „VEGA“ from the active selection (using set difference) by clicking on the respective bar or sector while holding a special key modifier (Fig. 3.7). Alternatively, using other key modifiers, it is also possible to restrict the selection to elements that belong to set „VEGA“ (using set intersection), or to add the elements of this set to the selection (using set union). The visualization interactively updates the list of selected elements and the fractions highlighted in the bars, as well as the textual description of the query. This allows to visually guide the creation of the queries, by showing which elements are affected by a new filter *a priori*.

Figure 3.7: Defining visual queries on the elements interactively.

Interaction is also vital to steer the automated analysis of the data (Sect. 3.2), based on the insights gained by the visualization. For example, the data can be filtered and re-aggregated based on how they are distributed in the initial visualization. In Fig. 3.6, elements of degree ≥ 5 are filtered out, as these were correctly retrieved by the majority of the depicted classifiers. This aims to focus on classifiers that often succeed together, while the majority of the other classifiers fail. Interaction also, allows examining and comparing different association and similarity measures, to find ones that are suited for a given data set. Finally, interaction is vital to analyze if a subset of the data correlates with certain attributes (Sect. 3.2.5), by iteratively selecting this subset and exploring its element attributes in other views.

3.5 Visual-Analytics Paradigm - Revisited

The previews sections illustrate the three components of the proposed *VA* approach: automated analysis, visualization, and interactive exploration. These components facilitate a *VA* paradigm, following Keim's *VA* Mantra [79, p. 82]: "*Analyse first – show the important – zoom, filter and analyse further – details on demand*":

- **Analyse first:** The analytical methods presented in Sect. 3.2 are applied to the data first. This reduces a large $n \times m$ homogeneous tables into $m \times b$ bins that summarize how the n data points are related to the dimensions $A_1 \cdots A_m$ and to $m \times m$ links that summarize how the dimensions are related to each other. Additionally, the automated analysis ranks the data features $F_1 \ldots F_p$ according to how their distributions differ between certain groups of data (e.g. selected vs. unselected).

- **Show the important:** The wheel visualization (Sect. 3.3) shows the important information computed by the automated analysis. The bars reveal the elements that have high or low association with each homogeneous dimension. The links emphasizes strong relations by means of thickness. Additional visualizations, such as box plots, show the most relevant data features for certain analysis context such as features that enable strong separation between certain groups of data.

- **Zoom, filter and analyze further:** The interactive exploration environment (Sect. 3.4) enables highlighting or filtering out certain elements based on their associations and attributes. For example it is possible to filter out set-exclusive elements interactively when they form the majority of set elements, to focus on elements shared between the sets. Other examples of how automated analysis can be applied interactively are given at the end of Sect. 3.4.

- **Details on demand:** The interactive exploration environment (Sect. 3.4) is composed of multiple views that allow exploring the data at multiple levels of detail. Elements that are aggregated in certain bars or links can be explored individually in tabular lists. Also, the attributes of selected elements can be explored in additional views using suitable visualizations such as bar charts, histograms, and box plots.

Part II explains in detail how this *VA* paradigm is applied to analyze three classes of homogeneous data: set-typed data, classification data, and categorical data.

3.6 Comparison with Related Work

The proposed *VA* approach enables new insights in and analysis possibilities of homogeneous data beyond the state of the art. In the following, I compare the presented approach with the generic visualization methods presented in Chapter 2, depending on what is represented in the visualization and by scalability. Part II compares the approach with other techniques dedicated to visualize each of the respective classes of homogeneous data.

3.6.1 Comparison by What is Represented

Scatterplot matrices (SPLOMs) [56] provide details about how two dimensions correlate with each other, by showing how individual data points are scattered with respect to each pair of dimensions. Parallel coordinates plots (PCPs) [69] show this information between adjacent dimensions by means of connecting lines. Biplots [50] show this information implicitly, by means of distances between points representing projected dimensions on a 2D plane. The *table lens* [120] shows this information between one dimension that defines the table sorting, and the other dimensions. The proposed analytical methods summarize these relations between all the dimensions by means of similarity measures that are visualized as links in the wheel metaphor. This provides a compact overview of these relations, and additional information about groups of dimensions that are closely related together. Details about certain relations can be explored on demand, for example, in scatter plots. Fig.3.8 shows the same data depicted in Chapter 2 using the generic multi-dimensional visualization techniques mentioned above.

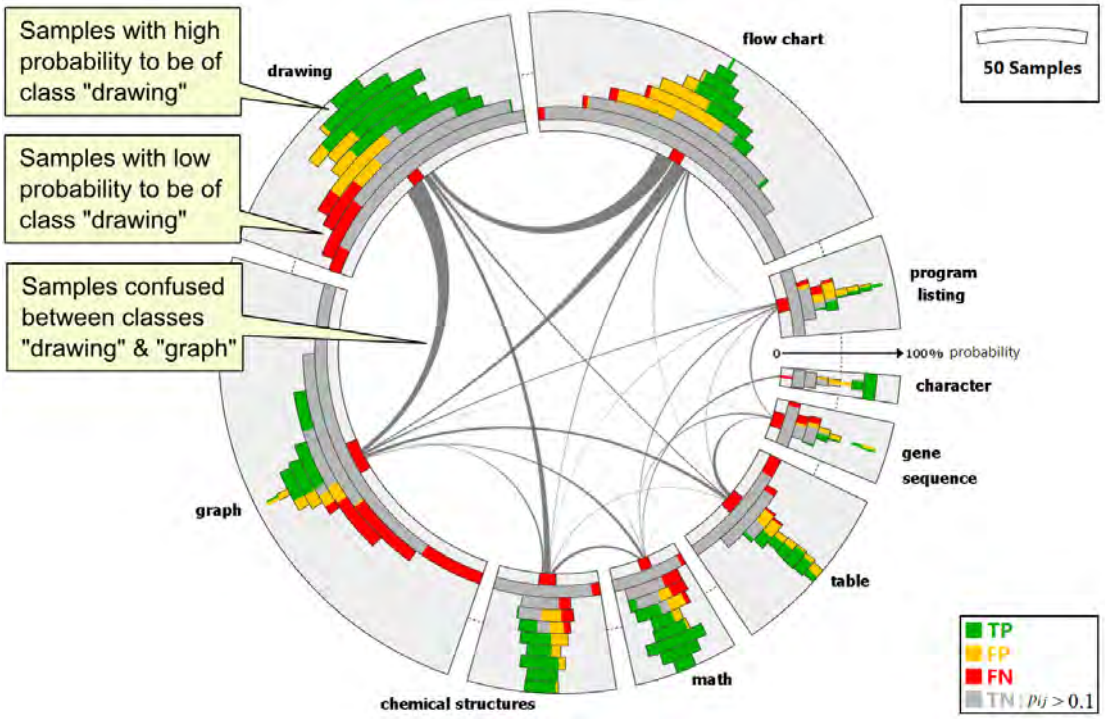


Figure 3.8: The wheel visualization showing the homogeneous data depicted in Fig. 1.1. Chapter B provides more information about the visualization.

In addition to the relation between dimensions, the proposed approach also computes aggregated association information of elements with dimensions, and visualizes this information as histograms. In contrast to the *table lens* [120], these histograms are depicted individually for each dimension, providing an overview of these associations for all dimensions at once.

Heat maps and matrices [56] offer a competing design to the node-link metaphor used to show the relations between the dimensions. Ghoniem et al. [51] compare between matrices and node-link diagrams and provide guidance on which metaphor is suited for which tasks and data complexity. It is easier to visually follow the link ends to infer the involved pairs of dimensions, than to follow separate rows and columns to infer the dimensions involved in the cells. However, this holds only as long as the number of links is small, since otherwise, the visual clutter will imbed visual navigation. The total number of symmetric binary relations between m dimensions is $\binom{m}{2}$. When the number of significant relation varies relatively in small range $[m..2 * m]$, the node-links allow fast detection of these relations as thick links (assuming the remaining links are thin and not relevant for the analysis). Heat maps are better suited to gain overview of the relations when the number of significant relations is large. Cell color is also more suited to reveal small differences than link thickness, as the latter varies in a small range. Finally, matrices scale better in the number of dimensions (e.g., m = 200) than node-link diagrams. Nevertheless, the wheel metaphor has its advantages, when the number of dimensions is small $m \leq 30$. The histograms showing aggregated associations of the entities with the dimensions enable to visually relate this information with the dimension relations visualized via the links. Including this information in the matrix design is cumbersome, as its layout dictates how the space is divided, leaving limited space to show information at the boundaries. Fig. 3.9a shows how placing the histograms in a matrix layout limits their resolution and utility, even when they are scaled differently (which prevents direct comparison of their bars). Assigning visual primacy to the histograms enables better utilization of the space, even when they are scaled uniformly (Fig. 3.9b).

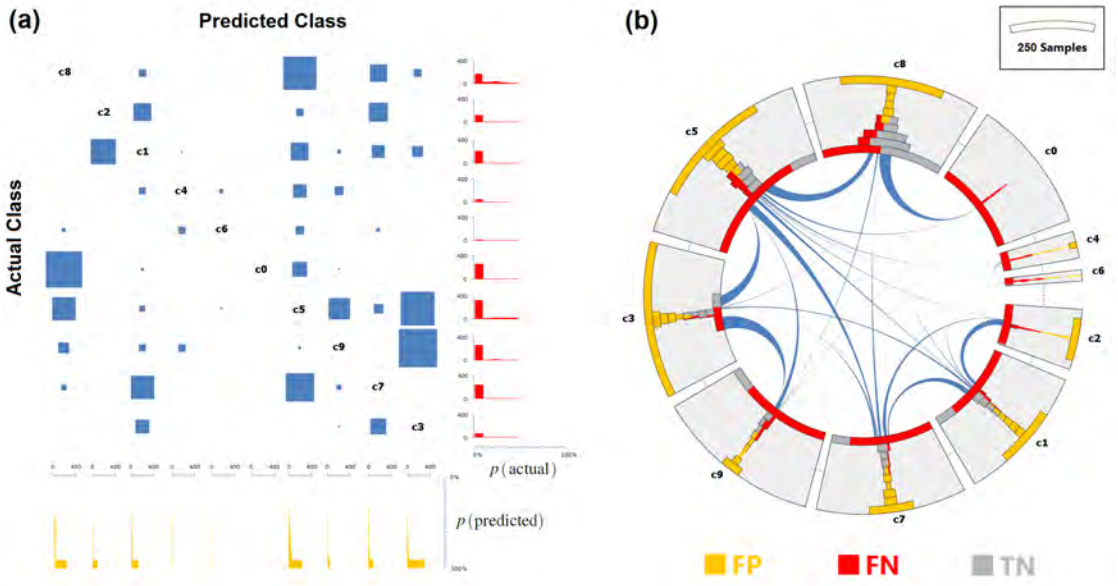


Figure 3.9: Comparison between a matrix-based design and the wheel design showing the same data. Chapter B provides more information about the depicted data.

3.6.2 Scalability

There are three aspects of *visual scalability* when visualizing homogeneous data:

- **Scalability in the number of dimensions:** The wheel metaphor can visualize about $20 - 40$ dimensions as sectors in a radial layout. With a larger number of sectors, the space assigned to the histograms becomes smaller, which undermines their value. Also, the crossings between the links and the associated visual clutter increases. The actual limit depends on the given dataset, and whether the histograms are divided as in Fig. 3.4a.

  In some datasets, the number of actual elements included in each histogram varies strongly, making few histograms occupy the majority of the space, while the remaining histograms are depicted at lower resolution. One way to handle such cases is to assign equal spaces to the sectors and to scale the histograms differently to fit in the sectors. While this hinders absolute comparison of the length between different histogram bars, it still allows comparing the overall distributions, by comparing the histograms *shapes*.

  One way to handle a larger number of dimensions is to include only a subset that fits in the visual limits, as explained in Sect. 3.2.2. The interactive exploration environment allows selecting the dimensions to be included in the wheel view interactively, based on information depicted in the overview views (Fig. 3.6). The remaining dimensions can either be excluded or aggregated into one special group with a suitable label (e.g., "others"). For example, dimensions that represent sets can be merged using set union. In case the dimensions exhibit a hierarchical aggregation, the analysis can be restricted to the aggregation levels that fits into the visual limits, with certain groups expanded on demand.

- **Scalability with the number of data points $E = \{e_1, ..e_n\}$:** The wheel metaphor exhibits high scalability with the number of entities n, thanks to the aggregations described in Sect. 3.2.2. Only a total of $m \times k$ histogram bins are used to show aggregated entities, instead of showing all the n entities individually (where k is the number of bins used for binning the association values). This allows analyzing datasets where n is in the order of tens to hundreds of thousands, as illustrated in Part II.

- **Scalability with the number of additional heterogeneous attributes:** The wheel metaphor is limited in its ability to show additional attributes of homogeneous data. It can use color to show information of how one attribute is distributed in the entities aggregated in the histogram bars. In case the attribute is categorical, only a few categories ≤ 7 can be displayed by dividing the bars into sections. In case the attribute is numerical, only a summary statistic about its values among the aggregated entities can be displayed, such as the average or the median.

  The detail views compensate for this limitations, by visualizing the attributes at a higher resolution and level of detail using suited visualizations such as bar charts, histograms. Multiple attributes can be visualized at once. Also, compact visualizations such as boxplots can be used to compare the distribution of multiple attributes (e.g., 20 attributes) between different subgroups of the data (Fig. 3.5).

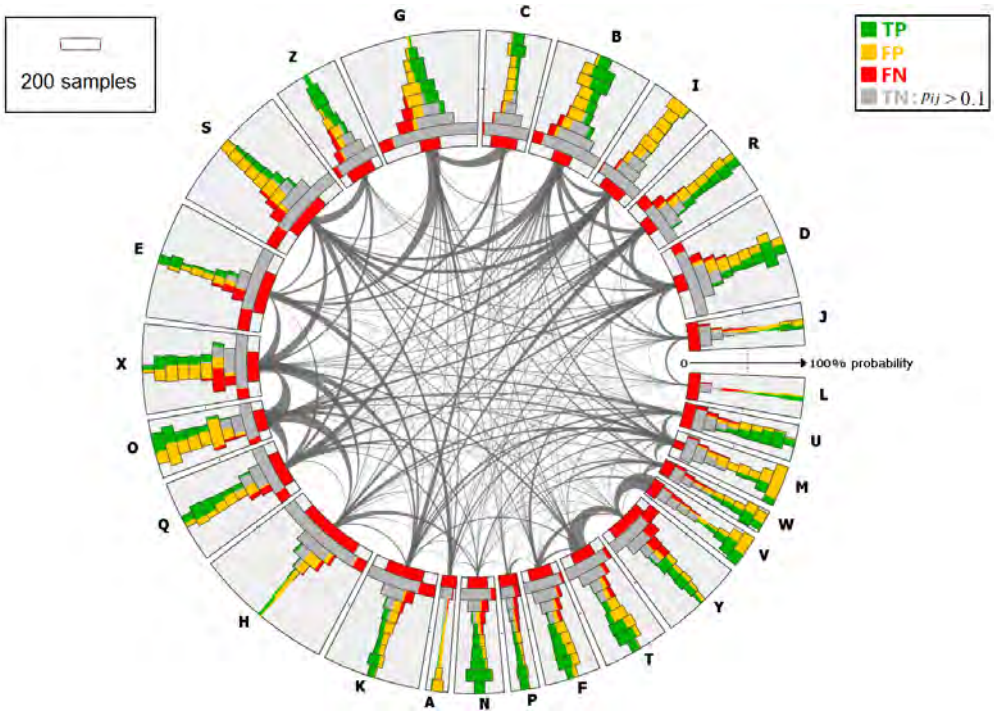


Figure 3.10: Scalability limitations of the wheel metaphor. The visualization shows classification results of Latin letter in 26 classes (see Chapter B for more details about this application).

The wheel metaphor is hence more scalable with the number of data points than SPLOMs [56] or biplots [50] that depict these points individually. It is also more scalable than PCPs [69] both with the number and dimensions and data points. Compared with the *table lens* [120], the wheel metaphor exhibits similar scalability limits, given that multiple rows can be aggregated in one row in the table lens. Compared with a heat map showing raw data values (Sect. 2.1.5), the wheel metaphor is significantly more scalable in the number of data points, if the heat map does not aggregate the row items. Aggregating the rows of a heat map usually leads to information loss, and the result heavily depends on the row order.

With respect to *computational scalability*, the presented methods allow fast automated analysis of large volumes of homogeneous data, encompassing tens of homogeneous attributes, tens of additional heterogeneous attributes, and hundreds of thousands of data points. This is because the proposed association and similarity measures (Sect. 3.2) are fast to compute in a straightforward way based on the given equations, and do not involve complex computations. Furthermore, many of these values can be pre-computed at the beginning of the analysis, without a need to be re-computed upon interaction. The correlations with the additional attributes, and the significance analysis of their deviations between multiple groups can also be computed efficiently for the selected entities during the analysis. Finally, producing the visual layout from aggregated data is computationally inexpensive. Besides simple geometric computations, an $O(m^2)$ greedy algorithm is employed to compute the sector ordering, as explained in Sect. A.3.1.

3.7 Evaluation and Limitations

This section summarizes the evaluation results of the proposed approach, conducted using different methods. I first explain how the proposed visual design was revised based on a pilot evaluation study of an earlier design. Then I explain the evaluation methods used for each application, and summarize the results. Finally, I report perceptual limitations of the proposed design that impact the accuracy of perceiving the depicted information. Part II provides more details about evaluation results for each of the three design instantiations.

3.7.1 Early Item-based Visual Design

The wheel metaphor proposed in Sect. 3.3 is based on an earlier design developed by the author [6]. This design also uses a radial layout to show the homogeneous attributes as sectors, and links to show relations between them. However, the data points were depicted as individual dots inside these sectors, without aggregation into histograms (Fig. C.3c). The radial location of a dot encodes information about the association of the respective entity and homogeneous attribute. The angular location encodes no information and is used to alleviate overplotting the dots.

The early design was evaluated via a qualitative pilot user study to assess its utility and usability [114]. The study was conducted by the Institute for Design & Assessment of Technology at the Vienna University of Technology. The data used for the evaluation was a 94×9 contingency table, extracted from a medical survey. The rows represent 94 survey questions, the columns represent answer agreements between teachers and parents. The table was visualized using the early version of the contingency wheel [6], with many features disabled to simplify the user interface [114]. The participants were ten computer scientists, five of which were visualization experts. After a tutorial part, the participants were asked to select and filter certain data elements and to answer questions about the visualization and the data sets. A qualitative assessment of the visualization was conducted based on the participants' responses.

As the study shows, most participants state that the radial layout provides a good first overview of the entire data, without a need for scrolling, and with the ability to select certain items for investigation. Additionally, the participants appreciated filtering and the possibility to explore the underlying data in a linked tabular view. On the other hand, the participants found difficulty in interpreting the links and the locations of the dots in the sectors (as the angular location is arbitrary). They also found the overall design to be very complex to understand.

3.7.2 Revised Visual Design

The wheel metaphor as presented in Sect. 3.3 was designed by revising the original design based on the results of the pilot user study. The main difference lies in aggregating the entities into histograms, instead of showing them as individual dots. This avoids confusion about what the angular locations of the dots represent, as the length of the histogram bars encodes information about their quantity. Informal follow-feedback from four visualization experts encourages this decision. These experts find histograms to be a good way to represent the information in aggregated form since histograms are easier to interpret than individual dots, besides avoiding several issues with overplotting and layout artifacts caused by the dots.

The revised wheel metaphor is applied in this thesis to visualize three different classes of homogeneous data, where different forms of evaluations were used for each class:

Classification data: This application targets machine learning users who aim to understand and compare the behavior of their classification algorithms on a given dataset. As explained in Chapter B, the visual design was adjusted to fit the nature of classification problem, and was refined iteratively based on feedback from machine learning experts. Chapter B provides usage scenarios with real datasets as well as details about expert feedback on the visualization design. Seven machine-learning experts and practitioners provided feedback based on an interactive interactive prototype. Additionally, nine experts reviewers provided anonymous written feedback based on a manuscript that describe the work. The experts agree that the design provide a good overview of the data, and reveals interesting patterns that explain the classifier behavior. However, about half of them find the design complex and requires extensive training and explanation to understand. These experts consider the presented system useful for analyzing classification data if the metaphor is understood correctly.

Overlapping sets: Overlapping sets are the simplest class of homogeneous data addressed in this thesis. Chapter A provides usage scenarios using the wheel metaphor with real set-based data. Chapter 4 provides case studies with domain experts in gene analysis and survey analysis, as well as example applications in machine learning and information retrieval.

In a recent work, Rajjo [118] conducted a task-based quantitative evaluation of the wheel metaphor for visualizing overlapping sets. The MovieLens dataset [53] was used for the evaluation, encompassing membership information of $3,883$ movies into 17 genres, along with movie release date and average rating. 32 participants were university students having different backgrounds including computer science, engineering, management, and social sciences. After a 20-30 minutes tutorial part, the participants were asked to use the visualization to answer 53 questions about the data, as well as to provide qualitative feedback about their experience with the visualization. The questions were divided into three groups, according to a recent task taxonomy [8]: tasks related to element-set memberships, tasks related to set relations, and tasks related to element attributes. Example questions are: "Name a movie that belongs ONLY to Thriller genre", "How many movies belong to Musical and Children at the same time?", and "Name a genre that tends to have a low average rating". The questions in each group were also divided further into three levels of difficulty, based on the number of set operations or deaggregation levels needed to solve the tasks. The performance was measured in terms of the time needed to answer the questions and the correctness of the answers. Table 3.1 summarizes the average performance measurements for each of the task groups and difficulty levels.

Based on the evaluation results, Rajjo concludes that the visualization is effective at solving the evaluated tasks, given that the users were provided with sufficient training and examples (about 20-30 minutes). The qualitative user feedback he collected also confirms that the metaphor is straightforward to learn with all users agreeing that histograms were intuitive to perceive a groups of elements. Two users found the sectors difficult to perceive as sets, and three users found the links difficult to perceive as set overlaps. Refer to [118] for more details about the experiment design and the result analysis.

	Element-related Tasks		Set-related Tasks		Attribute-related Tasks	
Difficulty	**Correctness**	**Time**	**Correctness**	**Time**	**Correctness**	**Time**
Easy	100%	32.2	99.5	36.5	100%	57,4
Intermediate	99.5%	31.1	98.8%	40.8	100%	64.4
Hard	85.6%	58.9	86.5%	112.3	97.7%	90.7
Overall	94.7%	39.6	95.6%	59.1	99.3%	70.9

Table 3.1: Rajjo's evaluation results [118] of the Radial Sets visualization (Chapter. A). Correctness is computed as the rate of correct answers of questions in each task group and difficulty level. Time is computed as the average time in seconds needed to answer these questions. Refer to [118] for information about confidence intervals of the above values.

Categorical data / contingency tables: This application scenario targets the analysis of associations between the categories of two categorical attributes, measured as disproportionally high (or low) frequency for pairs of categories to appear together. Chapter C provides usage scenarios using the proposed visualization with public data sets on movies rated by users [53]. It demonstrates how the visualization helps in identifying strong associations between movie genres and users or between user occupations and movies. It also demonstrates how coloring helps in correlating these associations with additional attributes of the users and the movies.

The informal feedback collected by the author from more than ten visualization experts and practitioners [2] suggests that the visualization as proposed in Chapter C is rather complex to understand. It requires a lot of explanation and a strong understanding of the mathematical notion of categorical associations. The filtering and interaction possibilities proposed were also not straightforward to apply, particularly in understanding their effects. This is partly due to the inherent complexity of contingency data [114]. Based on these remarks, the depicted information needs to be simplified and more annotation and visual aids should be provided to facilitate understanding. Other visual representations are potentially better suited to represent association information than histograms. In particular colored bins or matrix cells are probably suited to show which categorical combinations are highly associated with each other, as demonstrated by Bernard et al. [15]. More work is needed to make such matrix-based representation as scalable as the histograms.

Based on the above findings, the wheel metaphor is best suited to visualize overlapping sets. This is mainly due to the simplicity of this data type and the intuitive concepts associated with sets and set algebra that are used for the visual mapping. The metaphor is also suited to visualize probabilistic classification data, given that (1) users have enough experience with the notion of such data and associated concepts such as class probabilities, class confusions, false positives, etc. (2) users are given sufficient training to understand the visual mapping and the interaction possibilities. Finally, the metaphor is less suited to show contingency data, as to the average users, the visualization and the underlying mathematical concepts might be too complex to understand.

[2] This feedback was received to a poster at the IEEE VisWeek 2012 conference in Seattle.

3.7.3 Perceptual Limitations

The radial layout of the wheel metaphor suffers from perceptual limitations that impact the accuracy of reading the histograms bars. The two main limitations are related to how the actual bar lengths are perceived, and how they can be compared against each other.

When the wheel sectors are created by dividing a ring charts, the histogram bars need to be bent accordingly to fit in these sectors and to have equal bar thicknesses. However, the innermost bar of a histogram is placed on a smaller circle than the outermost bar. This makes bars of identical length appear differently, as the innermost bar spans a larger angle than the outermost bar, making them appear longer (Fig. 3.11a). This effect is minimized when the bars are short, as short arcs are perceptually flattened [123]. It is also possible to manipulate the bar lengths to make equal bars appear of the same length.

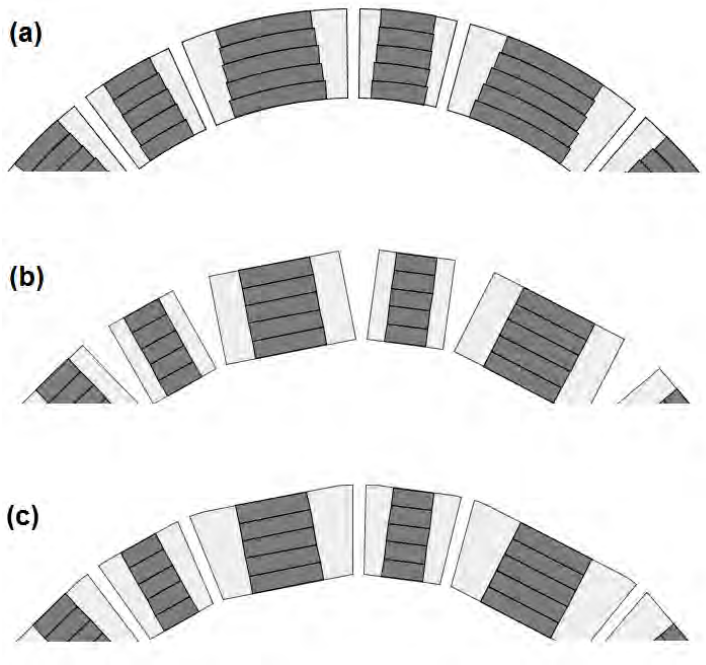


Figure 3.11: Perceptual issues with the wheel metaphor: (a) bending histogram bars causes bias in length perception, (b) the radial arrangement of rectangular sectors causes bias in shape perception, (c) closing the white gaps in (b) might lead to artificial 3D cues.

An alternative to bending the histogram bars is to flatten the sectors. In Fig. 3.11b, the sectors are depicted as rectangles with a radial arrangement. However, this causes irregular spacing between the sectors, and distorts the perception of rectangles that appear as trapezoids. A solution for these issues is proposed in Fig. 3.11c. Sectors are extended to fill the irregular spacing, leaving equally-sized gaps between the sectors. This however, adds artificial 3D cues to the sectors. Furthermore, flattening the sectors distorts the symmetry of the radial layout, especially when sectors vary significantly in size.

Another issue with the histograms is that the bars are centered in their sectors instead of having a baseline. The rationale behind this layout is to avoid artificial asymmetry across histograms in different sectors, and to make comparing their shapes easier. Moreover, the symmetry facilitates perceiving the histograms as figures or objects in their sectors following Gestalt laws [165]. This emphasizes the fact that the elements aggregated in the histogram are contained in the respective sectors which represent sets, classes, or categories.

Both the radial arrangement and the centered layout of the histograms impact the accuracy in perceiving and comparing the length of different bars. Therefore, the wheel metaphor is suited to provide an overview of item distributions in different sectors. Other views are needed to perform precise comparisons between or judgment about certain values. Nevertheless, the overview is useful to detect different distribution patterns, and to find gaps or extreme values in certain histograms.

CHAPTER 4

Example Applications

Chapter 3 presented the *VA* approach proposed in this thesis for analyzing homogenous data. Part II explains in detail how this approach is applied to three classes of homogeneous data, along with example applications. The next sections describe two actual applications of the approach in case studies with domain experts, along with two additional example applications. These studies and examples involve analyzing set-typed data, that are the most common data class addressed in this thesis. Probabilistic classification data stem for a specific application example described in detail in Chapter B. Contingency data are more restricted in applications than set-typed data due to their complexity, as discussed in Sect. 3.7.

4.1 Analyzing Genetic Data

This application is concerned with the exploration of genetic data among a diverse set of individuals. This problem encompass the following aspects, following the design triangle [104] for creating user-centered design of VA solutions:

- **Users:** A collaborator specialized in genetics and working as post-doctoral research fellow in the school of medicine at Stanford University. He contacted the author after he read about the set visualization technique proposed in Chapter A, and found it potentially suited for his application.
- **Data:** The collaborator provided a dataset that encompasses $9,651$ differentially-expressed gene transcripts for 14 individuals. Each of the individuals possesses a set of transcripts. These sets are intersecting, since the same transcript can be expressed in multiple individuals. The set elements are the transcripts. These elements have attributes such as RNA expression, ribosome occupancy, translation efficiency, and protein level.
- **Tasks:** The collaborator wanted to analyze the intersection relations between these gene sets. He was also interested in analyzing how an attribute is distributed in different sets and set intersections.

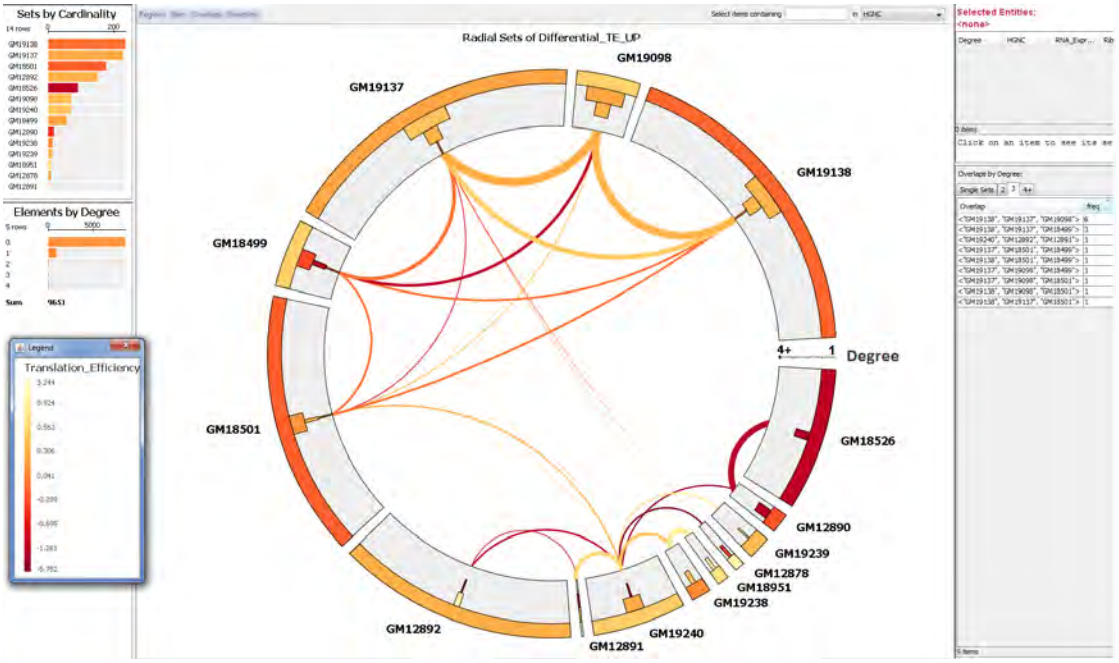


Figure 4.1: Visualization of gene transcripts in 14 individuals. The transcripts are colored by their translation efficiency.

The gene transcripts $\times$ individuals data can be visualized using Radial Sets, an instantiation of the proposed VA approach for visualizing overlapping sets. Fig. 4.1 shows an example of this visualization. The sectors represent the individuals, with histograms representing their gene transcripts. These histograms provide an overview of how many genes each individual shares with how many other individuals, as requested by our collaborator. An arc between two individuals represents genes shared between them. The bar color is mapped from „translation efficiency“ of the gene transcripts and shows the median of this attribute among the transcripts aggregated in the bar. The visualization enables flexible selection and filtering of certain gene transcripts, such as ones exclusive to a specific individual, or shared between two specific individuals. The selected gene transcripts are accessible in a tabular list, along with their attributes. This list can be exported for further analysis in other programs, as requested by our collaborator. Using these features, our collaborator was able to gain the following insight in the data:

- Similarities between individuals roughly but not perfectly recapitulate the population that they are coming from.
- The individual relationships are highly consistent globally between RNA expression and Ribosome Occupancy.
- Overall, the type of genes that are differentially expressed fall into specific categories when comparing RNA Expression vs. Ribosome Occupancy.

The collaborator commented on his experience with the visualization as follows:

„I also think that the true power of the approach is to enable other researchers to look at their favorite sets of genes or GO[1] *categories. Therefore, I think it will be very useful to include the GO categories“.*

More details about this application can be found in the following upcoming article: Cenik et al: „Integrative analysis of human variation in RNA expression, translation efficiency and protein levels“, 2014 *under review*.

4.2 Survey Analysis

Survey data is rich of information that enable analyzing the relation between the survey answers and relating the subjects' attributes with these answers. Binary yes/no survey questions define sets over the survey subjects. Each of these sets contain those subjects who answered the respective question with „Yes“. Another case where sets arise in survey data are questions that allow crossing multi-choices as answers, such as the languages an applicant masters. Each of these choices define a set over the survey subjects containing subjects that crossed this choice (possibly besides other choices).

This application encompasses the following aspects:

- **Users:** A collaborator specialized in communication sciences and working as Assistant Professor (affiliation removed to prevent identification). He contacted the author after finding the visual metaphor presented in this thesis potentially suited for his application.

- **Data:** The collaborator provided a dataset that encompasses subject answers to 10 yes/no questions.

- **Tasks:** The collaborator was interested in analyzing how the answers to certain questions influence the answers to other questions. He was also interested in analyzing if an attribute of the subjects influence their answers.

The survey data can be visualized using Radial Sets, an instantiation of the proposed *VA* approach for visualizing overlapping sets (Chapter A). Fig. 4.1 shows an example of this visualization. Each question represents a set (question names anonymized). The histogram in each set represents subjects who answered the respective question by "Yes". These subjects are grouped into bars according to how many questions they answered by "Yes". The bars are colored by an attribute of the subject.

The collaborator appreciated the insight provided by the visualization, in particular, to find questions that were more often answered by 'Yes" together. He also appreciated the interaction possibilities to select certain subjects based on their answers and to explore their attributes.

[1] Gene ontology.

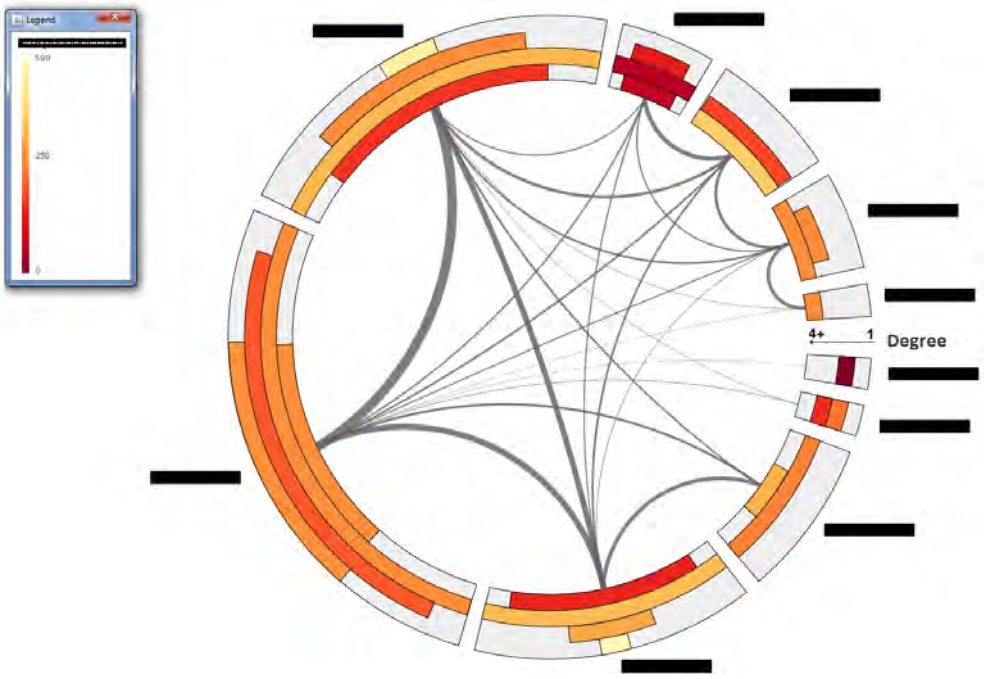


Figure 4.2: Visualization of survey answers, represented by sectors.

4.3 Comparing Multiple Classification Algorithms

Classification is a central problem in machine learning, involving predicting the class (out of many classes) a given sample belongs to. When the ground truth is available, it is possible to compare the predicted classes with the actual classes of the samples, to assess the classification performance. Multiple classification algorithms can be compared against each other by means of overall performance. However, even when algorithm A has higher performance than algorithm B, there are samples that algorithm A classifies better than algorithm B. This leads to the following analysis problem:

- **Users:** Machine learning experts.
- **Data:** Classification results of a set of samples by multiple algorithms. These algorithms can be modeled as sets over the classification samples, with each set containing the samples that are correctly classified by the respective algorithm. The samples can have further attributes such as their actual classes or the data features used for classification.
- **Tasks:** Comparing multiple classification algorithms in detail, beyond comparison of overall classification rates. This involves finding which samples are correctly classified by certain algorithms and not by other algorithms, and what the actual classes and the data features of these samples are.

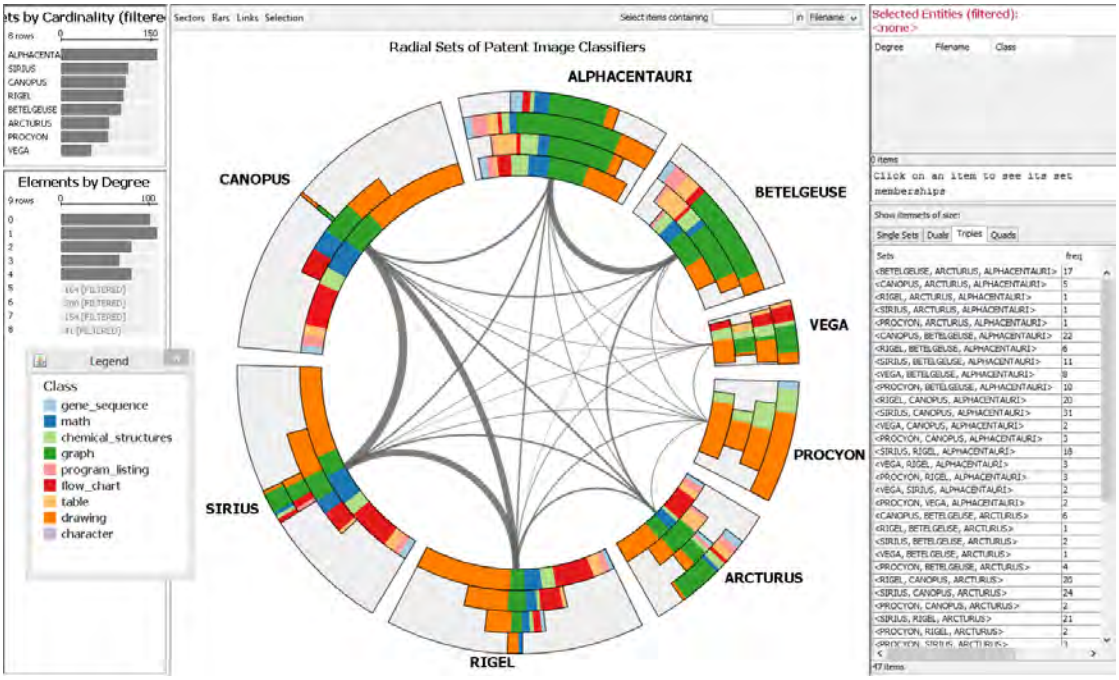


Figure 4.3: Comparing multiple classification algorithms, represented by sectors. Each sector contains samples classified correctly by the respective algorithm. The samples are grouped and filtered by their degrees and are colored by their actual classes.

Fig. 4.3 shows a visualization of classification data computed by eight classifiers. These classifiers were developed by Joanneum research center during the CLEF-IP 2011 classification evaluation campaign[2]. These classifiers are depicted as sectors. The dataset being classified comprises 1,000 patent images that are classified into nine classes as described in the color legend. These samples are depicted as histograms and are colored by their actual classes. The histograms reveal that all samples correctly classified by „BETELGEUSE“ were also correctly classified by at least one other classifier. This is revealed by the outermost bin, where set-exclusive elements are depicted, being empty. To focus the analysis on samples that were more specific to their classifiers, samples that were recognized by the majority of classifiers (five to eight) are filtered out. This shows that the classifiers „ALPHANUMERIC“ and „BETELGEUSE“ excel at classifying „graph“ images, compared with the other classifiers. On the other hand „PROCYON“ excels mainly in classifying images of classes „drawing“ and „chemical structures“. There is also a visible cluster of three classifiers („CANOPUS“ „SIRUS“ and „RIGEL“) that frequently succeed on the same samples while the other classifiers fail. It is also possible to color the samples by their attributes, to see if certain classifiers perform better than most of the others for certain attribute values.

[2] `http://www.ir-facility.org/call-for-participation1` - accessed Aug 1, 2014

4.4 Supporting Faceted Search

Faceted search [153] is an information-retrieval technique which assumes that the data items being searched can be multi-classified into facets. This classification paradigm assigns one more facets to the data items, based on their properties. The items are retrieved based on these facets, where users can define which facets they want to include in the search results. Very often, few or no search results are returned that satisfy all search facets, which causes the user to iteratively refine her search. This process is not only tedious and time consuming, but it might cause the user to miss relevant results that partially satisfy the search constraints. The user needs to gain an overview of how many data items satisfy each combination of her search facets. This leads to the following analysis problem:

- **Users:** information seekers who search data items by their facets.
- **Data:** multi-faceted data items, where one data item might satisfy multiple search facets. Each facet can be defined as a set containing data items that satisfy it.
- **Tasks:** (1) gaining overview of how many data items satisfy different combinations of search facets, and (2) defining search queries iteratively based on which facets can still be satisfied among items that satisfy selected facets.

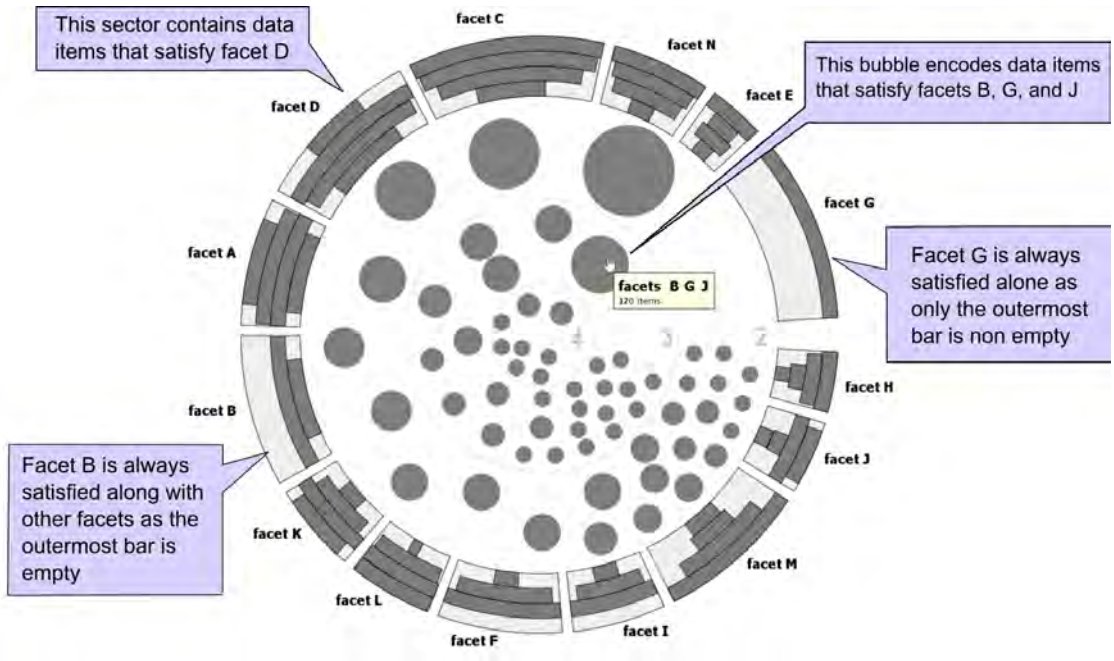


Figure 4.4: Visualization of multi-faceted search data. Bubbles are depicted instead of links to indicates all facet combinations. This demonstrates the extensibility of the wheel metaphor.

Fig. 4.4 shows the proposed set visualization for multi-faceted data. The sectors represent the search facets, with a histogram in each sector showing the data items that satisfy the respective facet. This histogram distinguishes between the items in each facet by how many other facets they satisfy. Items that satisfy only one facet are included in the outermost histogram bars next to the facet label. Items that belong to two facets are included in the next histogram bar, and so on. Hence, the histogram bars closer to the central area contain items that belong to more facets. It is possible to see immediately from the visualization, if items satisfying a certain search facet tend to satisfy other facets or not. For example, items satisfying 'facet G' in Fig. 4.4 cannot satisfy other search facets, whereas all items satisfying 'facet B satisfy two or three other facets.

Instead of showing links between pairs of facets, the visualization shows how the items are distributed into all *facet combinations* by means of bubbles in the inner area. The bubbles are organized into concentric rings, with the outermost ring containing combinations between pairs of facets, the next ring contains combinations between three facets, and the innermost ring contains combinations between four facets. The size of a bubble encodes the number of data items that satisfy the respective facet combination. The bubbles are sorted by size in the respective ring. To avoid visual clutter, the visualization initially does not indicate which facets are involved in a specific bubble. This information can be obtained on demand by hovering the mouse over a bubble. The respective facet combination is shown in a tooltip and can additionally be highlighted by means of visual links between the bubble and the respective sectors. It is possible to include rings starting for a specific number of facets, to focus on items that satisfy as many facets as possible.

Interaction allows selecting certain items in the visualization to explore these items in detail. The exploration environment allows showing selected search items in a tabular lists, along with their attributes. The visualization helps the user in interactively defining her faceted search query. For example, the user can start by selecting 'English' as an important skill when searching for a job candidate. This highlights matching portions of all histograms for all facets. Based on this visual guidance, the user can iteratively check and decide which facets can still be satisfied, and then narrow the search by combining selections using 'AND' operator.

This application demonstrates the extensibility of the wheel metaphor to address the specific requirements of certain applications. It was possible to replace the links between pairs of sectors with bubbles that represent the intersection between multiple search facets, and to organize these bubbles by the number of facets involved in them. Chapter. A provides more details about this extension and how it can be applied to visualize information specific to set-typed data. It also demonstrates how hyperlinks of degree 3 or 4 can be shown between the bubbles and the sectors. Further details about the applications of this visualization to multi-faceted search are available in [9].

CHAPTER 5

Conclusion

I conclude this thesis by summarizing the main contributions and limitations and by revisiting the hypotheses and research questions. I also reflect on the lessons learned from the conducted research, and discuss possible research directions for future work in the research topic addressed by this thesis.

5.1 Summary of Contributions and Limitations

This thesis contributes novel *VA* methods for analyzing three classes of data:

- Element-set memberships (set-typed data).
- Probabilistic classification data.
- Categorical data / contingency tables.

It proposes a unified *VA* framework for these three classes of data by modeling them as instances of a broader class of *homogeneous data*. This framework encompasses:

- Automated analysis methods to abstract the data.
- Novel visualization methods to show important relations in the data.
- An interactive environment to explore the data at multiple levels of abstraction.

The thesis identifies the characteristics of homogeneous data, and the new analysis and visualization opportunities enabled by the homogeneity of the data. Homogeneous data encompass multiple dimensions that have the same nature, as a special case of multidimensional data. The thesis demonstrates how the unified *VA* framework supports various analysis tasks of the data classes listed above. An evaluation is conducted by means of user studies and use cases.

Several applications of the proposed *VA* approach are demonstrated with various datasets stemming from different domains. Examples include gene analysis, survey analysis, understanding classifier behavior and analyzing its performance, comparing classification algorithms, and supporting multi-faceted search.

The following list summarizes the major advantages of the proposed approach:

- Providing an overview of large homogeneous data encompassing thousands of data items.
- A high scalability with the number of data items, thanks to automated analysis methods that aggregate the data, and to the aggregation-based visual metaphor.
- Gaining new insights in data of the classes listed above beyond the state of the art.
- Enabling new analysis possibilities in homogeneous data, thanks to the multiple-level overview+detail exploration environment.

The following list summarizes the major disadvantages of the proposed approach:

- The complexity of the visual metaphor, which requires sufficient user training.
- The scalability with the number of dimensions is limited to about 20-40 dimensions.
- The visual metaphor exhibits low sensitivity to small differences in the data.
- The radial layout suffers from perceptual issues that impact the accuracy of reading and comparing the depicted values.

The evaluation results presented in Chapter. 3 suggest that the proposed methods are best suited for analyzing set-typed data, due to the simple notion of sets and set relations. In addition, the methods are useful for analyzing probabilistic classification data, given that the users are given sufficient training. Finally, the methods are less suited for analyzing categorical data summarized in contingency tables, as the proposed visual metaphor does not dierctly map to contingency tables.

5.2 Revisiting the Hypotheses and Research Questions

The thesis started with the following hypotheses and research questions (Sect. 1.2.4:

H1: ***VA methods provide new insights in and analysis possibilities of homogeneous data, that are difficult to gain or perform using automated data analysis method.***
This hypothesis is confirmed based on the case studies presented in Chapter 4 and the use cases presented in Part II. These cases demonstrate how the proposed VA methods enabled users and domain experts to understand and analyze their data and provided insight into interesting relations and patterns in the data, these experts were not aware about before. Sect. 3.4 demonstrates how interactive visualization is vital to steer automated analysis of homogenous data, which would be hard to perform without visual inspection.

H2: ***The structural similarities between different classes of homogeneous data enable reusing interactive visual analysis solutions across these classes.***
This hypothesis cannot be confirmed, and needs further study to confirm or reject. The unified *VA* approach proposed in this thesis was applied to three different classes of homogeneous data, as demonstrated in Part II. While this approach reveals new insights in each of these data classes, it requires sufficient learning in order to apply and to interpret the results correctly. The users

who participated in the user studies and the domain experts who led the case studies were confronted with one data class only. It remains unclear if users can adapt to use the same visual metaphor for analyzing a different class of data than the ones they learned first. Chapter. 3 technically demonstrates how the same abstraction can be applied to multiple classes, leading to a unified interpretation of the visualization (e.g. items that have high or low association with a dimension). However, when this abstraction is applied to a concrete data class, these associations have specific interpretations (e.g. set-exclusive vs. set-shared elements, or high vs. low class probability) that might confuse the users if they are combined in the same application.

Q1: ***How to analyze the association relations between a large number of row entities*** E ***and the*** m ***homogeneous column categories represented by the attributes*** $A_1 \dots A_m$***?***
Automated analysis is crucial to quantify these relations (Sect. 3.2.1), aggregate them into bins (Sect. 3.2.2) and to focus on the most relevant relations (Sect. 3.2.3). Aggregation-based visual representations (Sect. 3.3) together with an interactive exploration environment (Sect. 3.4) are crucial to explore these relations at multiple levels of abstraction, in a scalable way.

Q2: ***How to analyze the correlation between homogeneous attributes*** $A_1 \dots A_m$***?***
Automated analysis is crucial to quantify and summarize these correlations (Sect. 3.2.4). A radial node-link layout (Sect. 3.3) is suited to provide a compact overview of these relations and of groups of highly-correlated attributes. Interactive exploration (Sect. 3.4) is vital to analyze the summarized correlations between certain attributes in more detail.

Q3: ***How to analyze the influence (in both directions) of additional attributes*** F_k ***of the row entities on the row-column associations (Q1) and attribute correlations (Q2)?***
Automated analysis is crucial to compare the attribute distributions between different groups of data, to find significance differences between these distributions, and to measure the separability of these groups by the attribute (Sect. 3.2.5). Visualization is crucial to gain an overview of attribute distributions in different parts of the data (Sect. 3.3). Interactive exploration (Sect. 3.4) is crucial to select certain groups of data in order to analyze their attributes in detail.

5.3 Lessons Learned

The general lesson learned in this thesis is that *VA* is crucial to analyze large data sets that cannot be visualized in their entirety, and where automated analysis is very limited. More specifically, with respect to homogeneous data:

- Automated analysis is vital to reduce large volumes of data and to compute the most relevant information for interactive analysis, as stated by Keim in his *VA* mantra [79].

- Aggregation is an effective data reduction method coupled with an overview+detail exploration environment to retrieve individual elements that are aggregated in the overview.

- Homogeneous data lend themselves to new forms of aggregation and filtering, beyond the ones applicable to multi-dimensional data in general. Exploiting the homogeneity of the

data is key to computing effective aggregations and filtering that take into account the specific nature of the data and the tasks associated with them.

- Developing a new visual metaphor might result in a complex visualization that requires extensive training to understand and use effectively. Nevertheless, once understood, the visualization can lead to new insight and discoveries in the data.

- Set-typed data, encompassing element-set memberships along with element attributes, comprise a widely applicable and powerful data type. The raw data and the results of automated analysis are easy to understand, thanks to the simple notion of sets.

5.4 Future Work

Here I list open issues that were not addressed in this thesis, as well as potentially fruitful research directions for visualizing homogeneous data. Part II gives further pointers to research challenges specific to the respective class of homogeneous data.

Reducing the visual complexity One of the main limitation of the proposed *VA* approach is visual complexity. This can be addressed by investigating how to reduce the amount of information being depicted. Interactive filtering techniques could also offer a solution to this issue by allowing users to focus on certain relationships and attributes at a time.

Scalability with the number of dimensions The proposed visual metaphor can handle about 20 to 40 dimensions. One way to increase this limit is to investigate new forms of aggregations for homogeneous data, especially for set-typed and categorical data. These forms could further reduce the volume of the data. For example, hierarchical aggregation and exploration techniques could help to group certain dimensions and to expand them on demand.

Attribute hierarchies In some cases, the homogeneous attributes belong to a intrinsic hierarchy. For example, a classification can be defined at a coarse level which includes a small number of classes or at finer levels that specify the classes more narrowly. Dedicated methods should be investigated to support interactive exploration of such hierarchies and to analyze the relations and associations in the data at multiple levels of the hierarchy.

Interaction The main focus of this thesis was on developing computational methods and visual representations, where interaction was rather an afterthought in the proposed design. Effective and usable interactions add a lot of power to the visual representations and allow customizing the visualization to fit specific data characteristics and user needs. As mentioned in the previous points, new interactions can be investigated to reduce the visual complexity, to merge or aggregate certain parts of the data, and to support flexible and usable hierarchical exploration.

Evaluation Further evaluation of the proposed methods is needed to assess which features and parts of the design are useful, and which ones need to be revised. It is also important to perform comparative task-based evaluations of the proposed visualization against other visualization techniques and against alternative visual designs.

Alternative visual designs Several alternative visual designs can be proposed to show the information encoded in the wheel visualization. One alternative is to visually separate the entity associations (depicted via histograms) and the attribute relationships (depicted via links) and to use simpler representations for both pieces of information. Another alternative it to show relationships implicitly or on demand instead of showing all of them explicitly in one overview. This allows using a linear layout to avoid the complexity as well as the perceptual and scalability limitations of the radial layout. Finally, a matrix-based layout is a competing alternative that could reduce the visual complexity and provide a rich overview of the data.

Additional application examples Several application domains could profit from the methods proposed in this thesis. In particular, the field of computational biology and bioinformatics require analyzing large volumes of data that exhibit similar characteristics to the data classes investigated in Part II. Analyzing such datasets with the proposed methods can lead to better understanding of the data relations and might reveal hidden patterns and enable new discoveries in these fields. Multi-faceted search is also a promising application area that might profit further from analyzing set relations, as explained in Chapter. 4. Investigating such applications can also introduce refinements and alternatives to the proposed methods and help in assessing which features are more useful in actual applications than other features.

Other classes of homogeneous data There are other classes of data that can potentially profit from the *VA* paradigm proposed in this thesis. For example, a homogeneous data table might contain nominal values instead of binary values, probabilities, or frequencies. Additionally, fuzzy element-set memberships result in real and continuous values instead of discrete binary values. Identifying such classes and the tasks associated with them allows investigating the applicability of the proposed methods to analyze such data to gain new insight that was not possible before.

Part II

Papers

CHAPTER A

Radial Sets: Interactive Visual Analysis of Large Overlapping Sets

Appears in IEEE Transactions on Visualization and Computer Graphics, 19(12):2496-2505, 2013.

Abstract: In many applications, data tables contain multi-valued attributes that often store the memberships of the table entities to multiple sets such as which languages a person masters, which skills an applicant documents, or which features a product comes with. With a growing number of entities, the resulting element-set membership matrix becomes very rich of information about how these sets overlap. Many analysis tasks targeted at set-typed data are concerned with these overlaps as salient features of such data. This paper presents Radial Sets, a novel visual technique to analyze set memberships for a large number of elements. Our technique uses frequency-based representations to enable quickly finding and analyzing different kinds of overlaps between the sets, and relating these overlaps to other attributes of the table entities. Furthermore, it enables various interactions to select elements of interest, find out if they are over-represented in specific sets or overlaps, and if they exhibit a different distribution for a specific attribute compared to the rest of the elements. These interactions allow formulating highly-expressive visual queries on the elements in terms of their set memberships and attribute values. As we demonstrate via two usage scenarios, Radial Sets enable revealing and analyzing a multitude of overlapping patterns between large sets, beyond the limits of state-of-the-art techniques.

keywords: Multi-valued attributes, set-typed data, overlapping sets, visualization technique, scalability.

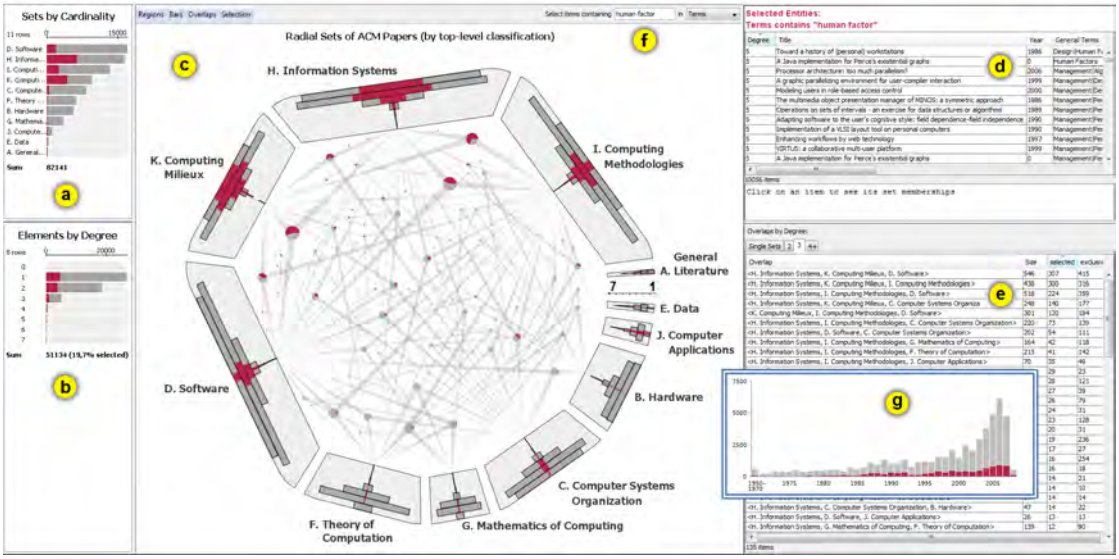


Figure A.1: The main interface of Radial Sets: (a) the sizes of the overlapping sets, (b) a histogram of the elements by degree, (c) the *Radial Sets view* showing $n > 50,000$ papers multi-classified into 11 ACM classes [1]; hyperedges of degree 3 are depicted to indicate overlaps between triples of sets; selected portions are highlighted in the bars and in the bubbles), (d) a list of $1,098$ selected elements and their attributes, along with a natural text describing the selection criteria, (e) the *overlap analysis view* showing details about overlaps classified by degree into different lists, (f) a search box to select elements containing a specific text, (g) a linked view showing the publication date distribution of all papers and of the ones in (d).

A.1 Introduction

Sets are one of the most fundamental concepts in mathematics. A set is a collection of unique objects, which are called elements of the set. Because of their simple and generic notion, sets are widely used in computer science to represent real-world concepts, query results, and the results of various algorithms. Compared to lists, sets ensure the uniqueness of their elements and impose no order on them. A set system comprises multiple sets defined over the same elements. Multiple set memberships are common in practice to represent both technical and real-world concepts. As an example, they can represent people memberships to different clubs, the markers a gene contains, or multiple tags or labels assigned manually or automatically to a set of entities. these memberships are usually stored in a database using either a multi-valued attribute or a group of Boolean attributes.

Sets defined over the same elements in a dataset potentially overlap. With a growing number of elements, these large overlapping sets contain a wealth of patterns that are worth to discover and analyze. Euler diagrams are the most common and natural way for depicting overlapping sets. However, they are inherently limited in terms of scalability.

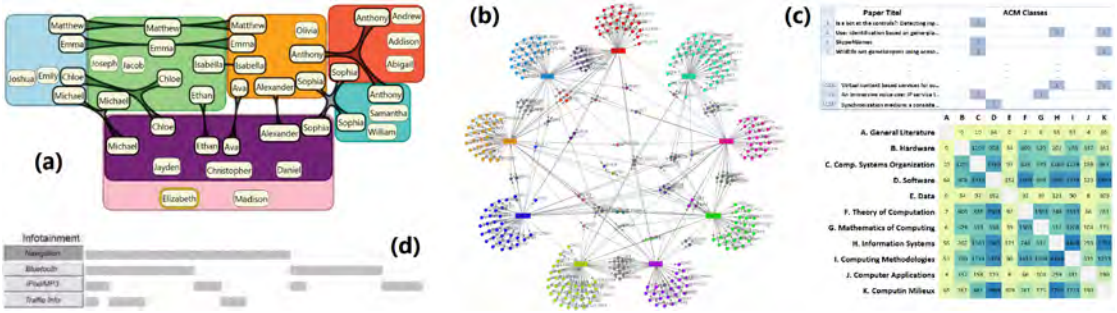


Figure A.2: Four techniques for visualizing element-set memberships: (a) untangled Euler diagrams [63] with duplications of elements that belong to multiple sets, (b) Anchored Maps [106] with sets represented as anchors on a circle and elements as free nodes, (c) two reorderable matrices [17] showing the element-set memberships and the set overlaps. (d) equal-height histograms [168] showing elements as bars in different rows,

In this paper we introduce a novel visualization technique for analyzing large overlapping sets. Our technique, called Radial Sets[1], shares several properties with state-of-the-art techniques proposed for the same purpose (Sect. B.2). It builds upon selected ideas from these techniques to improve both on readability and scalability, and to support advanced analysis and pattern-finding tasks for this kind of data. In particular, given set memberships of a large number of elements in about $m \leq 30$ sets, Radial Sets enable the following analysis tasks that are common for this kind of data [48,63,141]:

- **T1:** Analyze the distribution of elements in each set according to their degrees (the number of sets they belong to).
- **T2:** Find elements in a specific set that are exclusive to this set, or that belong to at least, at most, or exactly k other sets.
- **T3:** Analyze overlaps (intersections) between groups of k sets.
- **T4:** Analyze overlaps between pairs of sets: find which pairs of sets exhibit higher overlap than other pairs (related to T3).
- **T5:** Find elements that belong to a specific overlap.
- **T6:** Analyze how an attribute of the elements correlates with their memberships to the sets and the overlaps.
- **T7:** Analyze how set memberships and attribute values for a selected subset of elements differ from the rest of the elements.

The tasks **T1** and **T2** are concerned with element memberships in the sets. For example, if the sets are defined over products to represent the features they come with, a typical question about one feature is whether it tends to come exclusively, or along with one, two, or more other features. The overlap tasks **T3**, **T4**, and **T5** enable finding out which feature combinations

[1] A prototype implementation is available at www.radialsets.org

are more common among the products, and which products belong to these combinations. The attribute analysis tasks **T6** and **T7** answer questions like how the price of a product depends on its features and whether certain feature combinations are particularly cheap or particularly expensive.

As we show in Sect. A.3, the visual design of our technique is derived from the requirements of these tasks. It employs frequency-based representations of the set elements to support the memberships tasks **T1** and **T2** in a scalable way. Also, it dedicates a large portion of the screen space to emphasize the overlaps as first-order objects in the visualization, as required by the overlap tasks. Both the set elements and the overlaps are visualized using area-based representations. This supports using retinal variables [17] like color to show information about the elements, as required by the attribute-analysis tasks.

Sect. A.4 presents two usage scenarios of Radial Sets to demonstrate how they can be used to perform the tasks **T1,** $\cdots$**, T7** with large sets defined over thousands, to hundreds of thousands of elements. In Sect. A.5 we discuss the applicability and the limitations of Radials Sets, and outline possibilities for future work.

A.2 Related Work

Visualizing overlapping sets is a non-trivial problem due to the potentially large number of possible overlaps: there might be up to 2^m distinct intersections between m sets [159]. Each element lies in one of these intersections, based on its memberships to the different sets. Though many of these intersections is equal to the empty set in practice, the number of non-empty overlaps can still be large, even with a dozen sets. These overlaps are salient features of set data with many analysis tasks typically concerned with different kind of overlaps between the sets.

Some techniques for visualizing overlapping sets bypass the complexity problem by limiting the number of sets and overlaps that can be visualized at once. Other techniques avoid visualizing the overlaps explicitly and convey more abstract information about the set system instead. In the following, we categorize existing techniques based on the visual representations they use and discuss their scalability and which of the tasks listed in Sect. C.1 they support.

A.2.1 Euler Diagrams and Euler-like Diagrams

Euler diagrams [42] represent sets as closed regions in the plane, providing a very natural way to depict overlaps. However, they suffer from a severe limit: all possible overlaps can be depicted distinctively only with a small number of sets $m \leq 4$. Verroust and Viaud [160] showed that this limit can be increased to $m \leq 8$ by relaxing the conditions on the contours and by allowing holes in the regions.

Several techniques have been recently devised to automatically generate Euler-like diagrams. The methods of Flower et al. [46, 47] generate Euler diagrams in case of drawability. Rodgers et al. [124] and Simonetto et al. [142] presented techniques that generate an output even for undrawable instances by allowing disconnected regions. Both techniques can result in complex non-convex zones especially when the sets exhibit numerous overlaps. Henry Riche and Dwyer [63] proposed two variations to draw simplified rectangular Euler-like diagrams that also

represent individual elements. Their second variation, called DupED, does not depict the intersections between the sets explicitly. It rather creates separate rectangular regions for the sets, and duplicates the elements that belong to multiple sets. Multiple instances of the same element are then linked with hyperedges (figure A.2a). Recent work has focused on generating area-proportional Venn and Euler diagrams [28, 82, 166]. Such diagrams convey how large the overlaps are compared to each other without depicting the elements. However, generating these diagrams accurately is restricted to three sets.

Euler-like methods have also been employed to visualize set memberships over existing visualizations that determine the positions of the elements. BubbleSets [31], LineSets [3] and Kelp diagrams [36, 99] are examples of such methods with varying design goals and degree of compactness. Itoh et al. [70] proposed depicting the set memberships as colored glyphs inside the visual elements. Each set is hence denoted by disconnected regions linked only by having the same color.

In summary, methods based on Euler diagrams often impose severe limits on the number of sets, elements, and overlaps they can depict, and hence can only partially cope with the tasks **T2**, **T3**, **T4** and **T5**.

A.2.2 Node-link Diagrams

A set system of m sets $S_{1 \leq j \leq m}$ defined over n elements $e_{1 \leq i \leq n}$ can be modeled as a bipartite graph $G = (V1 \cup V2, E)$. The vertices of this graph are the elements $V1 = \{e_i : 1 \leq i \leq n\}$ and the sets $V2 = \{S_j : 1 \leq j \leq m\}$. The edges $E = \{(e_j, S_i) : e_j \in S_i\}$ are the membership relations between the elements and the sets. A variety of approaches were devised both for drawing [35, 107, 176] and for visualizing [106, 129] bipartite graph as node-link diagrams. Anchored Maps [106] place the vertices of one class as anchors on a circle. The vertices of the other class are placed as free nodes with links connecting each free node with the anchors it has edges with (figure A.2b). The position of these free nodes are determined by spring embedders.

A set system can be depicted as an Anchored Map by representing the sets as anchors and the elements as free nodes. This enables quickly finding which elements are exclusive to each set, and which elements are shared between multiple sets, partially solving the tasks **T2** and **T4**. However, with an increasing number of elements shared between multiple sets, the view becomes quickly cluttered making it difficult to recognize which elements belong to which overlap. This is an inherent limitation of node-link diagrams that restricts their applicability to a small number of elements.

Hypergraphs offer a more general way to model a set system with each set represented by a hyperedge that connects all element vertices in this set, or vice versa. The two general approaches to draw hypergraphs [94] roughly resemble Euler diagrams (subset standard) and node-link diagrams (edge standard).

A.2.3 Matrix-based Methods

A matrix can depict memberships of n elements represented as rows in m sets represented as columns (figure A.2c-top). Bertin described how reordering the rows and columns can simplify such matrices [17]. This ordering has a significant impact on the ability to find patterns in

the matrix, especially clusters of elements that exhibit similar patterns of memberships of the sets and vice versa [18, 167]. As the ordering problem is NP-complete [95], a large number of heuristics have been proposed for reordering matrices [92]. In addition, several interactive systems have been proposed to create and refine reorderable matrices for different purposes [64, 140, 146].

With a growing number of relations, the membership matrix outperforms node-link diagrams in several low-level reading tasks [51]. However, it falls short of solving tasks specific to set data. A separate matrix is needed to explicitly reveal the overlap between pairs of sets (task **T4**) as a heatmap (figure A.2c-bottom). Henry Riche et al. [65] augmented matrices with links that show additional relations between the rows or the columns (figure A.2c). Similar ideas can partially support **T4** in the membership matrix without the need for a separate matrix. Another problem with matrix representations is scalability: A large number of elements that belong to a smaller number of sets result in a skewed membership matrix. This is challenging for multi-level techniques that are usually designed for square matrices [41].

A.2.4 Frequency-based Methods

Node-link diagrams and memberships matrices offer item-based representations of overlapping sets that create a distinct visual item, like a node or a row for every element in the sets. In contrast to that, frequency-based representations aggregate multiple elements that belong to specific overlaps into a single visual item like a bar. This makes them potentially scalable in the number of elements they can depict.

Wittenburg proposed an extension to bargrams [169] to depict set-valued attributes [168]. The sets are represented as rows in the bargrams, sorted from the largest to the smallest. The horizontal dimension represents all the elements, sorted by their membership of the topmost set, then of the second topmost set, and so on. Bars are drawn in each row to depict the elements that belong to the corresponding set according to this sorting (figure A.2d). This reveals different overlaps between the sets, however, from the perspective of the larger sets which define the elements' order. A different ordering of the rows is needed to infer the overlap between the two bottommost set.

Set'o'grams [48] extend bar charts to visualize overlapping sets. Each set is represented by a bar of proportional size. This bar is divided into sections that represent the different degrees of elements in the respective set (figure A.3a). The degree of an element is equal the number of sets it belongs to. The sections are distinguished both from each other both by shading, and by assigning increasingly smaller widths to sections of higher degrees. Hence, it is possible to infer for each sets how many elements belong exclusively to it and how many of its elements belong to k other sets, solving exactly tasks **T1** and **T2**. Interaction by means of brushing can solve task **T5** but falls short of providing an overview of overlaps required for tasks **T3** and **T4**.

Our work extends the basic idea of Set'o'grams. It employs an alternative visual design that emphasizes the single sections in the bars and allows depicting different kinds of overlaps as we show next.

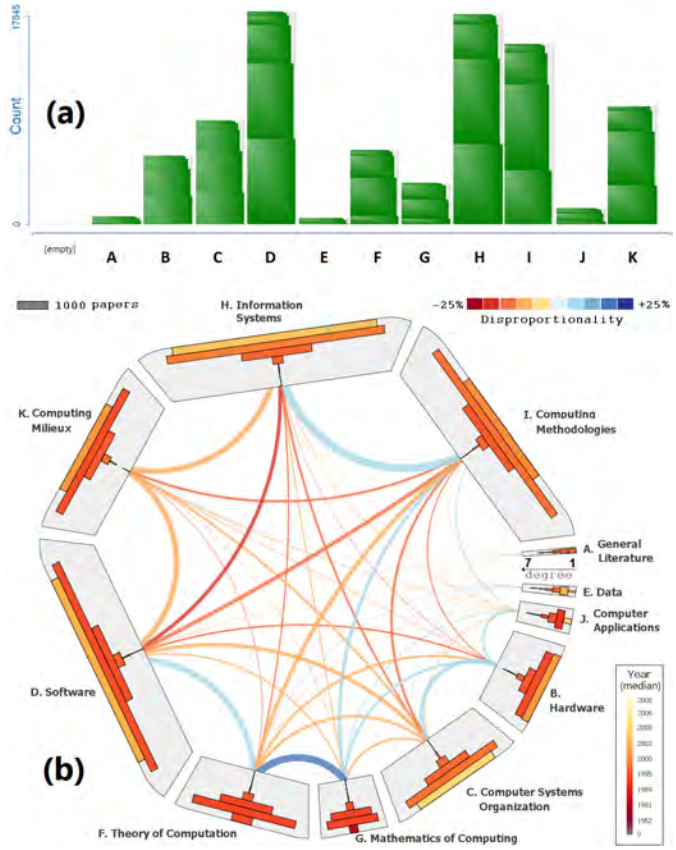


Figure A.3: (a) Set'o'grams [48] showing 11 overlapping sets as bars of proportional size, divided into groups of elements of equal degree, (b) Radial Sets showing the same data with overlaps between pairs of sets depicted as arcs. Ideally, only one color scale should be used.

A.3 Radial Sets

To enable a scalable visual analysis of large overlapping sets, Radial Sets employ frequency-based representations that aggregate the elements in the sets and in their overlaps. Also, multiple views depict the information at multiple levels of detail. The main view (Sect. A.3.1) shows both the distribution of elements in the sets and the overlaps between the sets. Additional views show both summary and detailed information about the elements and the overlaps (Sect. C.2.3). Together, these views enable an elaborate analysis of overlapping sets.

A.3.1 The Visual Metaphor

To visually encode overlapping sets, Radial Sets use three types of visual elements: (1) regions to represent the sets, (2) histograms inside the regions to represent the elements in each set, and (3) links between the regions to represent overlaps between the sets. Figure A.4 shows how four overlapping sets are represented as Radial Sets.

Visualizing the sets

Radial Sets represent the sets as uniformly-shaped non-overlapping regions. The regions are arranged radially on a circle. This arrangement aims mainly to ease the depiction of the overlaps between the sets as links inside this circle, and to emphasize them as the central part of the visualization. Moreover, it facilitates the interpretation of the histograms representing the elements in the individual regions as we explain in Sect. A.3.1.

Unlike Set'o'grams [48], the areas of the regions are not necessarily proportional to the sizes of the sets. A dedicated view in the user interface conveys these sizes more effectively via a bar chart (Sect. A.3.2). Depending on how the histograms are scaled, the regions can be either made of equal area or assigned different areas to fit the histograms. In the latter case, the regions are depicted as rounded parallelograms leaving equally-sized gaps between the regions. This alleviates visual artifacts and asymmetries caused by non-uniform gaps. However, the parallelograms might imply 3D cues to the regions, which impacts the accuracy of perceiving the bars insides these regions.

The use of distinct visual elements to represent the sets and the overlaps enables using simple shapes to depict the set regions. As discussed in Sect. A.2.1, a similar idea was employed by Henry Riche et al. to simplify Euler diagrams [63]. They argued that the use of convex and simple regions is a primary factor impacting readability, as shown by empirical results in Gestalt psychology [76]. We also duplicate the representations of elements that belong to multiple sets, like in the untangled Euler diagrams (figure A.2a). However, we aggregate these elements, and the overlaps they result in as we describe next.

Visualizing the elements

Like Set'o'grams [48], Radial Sets aggregate the elements of each set into groups according to their degrees. In a system of m sets $S_{1 \leq j \leq m}$ and n elements $E = \{e_i : 1 \leq i \leq n\}$, the degree of an element $e \in E$ is equal to the number of sets it belongs to:

$$degree(e) = |\{S_j : 1 \leq j \leq m \wedge e \in S_j\}| \tag{A.1}$$

The elements of each set S_j are aggregated via a histogram H_j of their degrees. Each histogram consists of $b = d$ bins with d denoting the largest number of sets that share at least one item:

$$d = max\{degree(e) : e \in E\} \tag{A.2}$$

Hence, the number of items in bin k of histogram H_j is:

$$h_{jk} = |\{e \in S_j : degree(e) = k\}| \tag{A.3}$$

It is possible to use a smaller number of bins b than d. In this case the last bin b aggregates elements having degrees equal to or higher than b:

$$h_{jb} = |\{e \in S_j : degree(e) >= b\}| \tag{A.4}$$

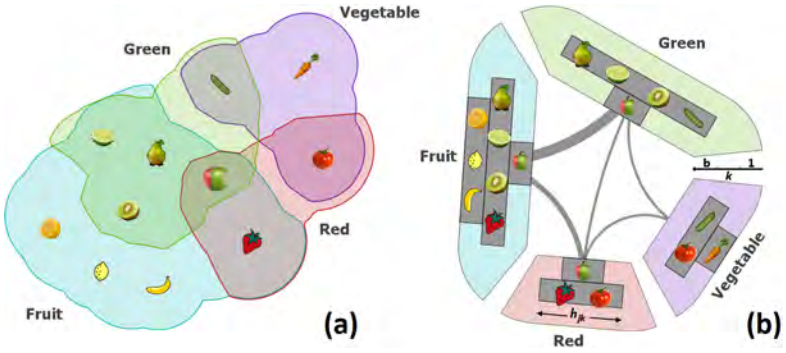


Figure A.4: (a) An Euler diagram (adapted from Wyatt [170]), (b) the equivalent representation in Radial Sets. The histograms in gray show a breakdown of the elements in each set by their degrees (Eqs. A.1, B.1). The arcs show overlaps between pairs of sets. The icons are for illustration only.

This aggregation limits the analysis to overlaps between $2, 3, \ldots$, till b-or-more sets. This is desirable since usually only few elements have high degrees. Aggregating them simplifies the visualization. The degree histogram retains access to these elements (Sect. C.2.3).

The histograms $H_{1 \leq j \leq m}$ are placed radially in the regions of their respective sets. The radial dimension encodes the elements' degrees k, with h_{j1} mapped to the outermost boundary of region S_j and h_{jb} mapped to the innermost boundary (figure A.4b). This intends to emphasize that the items in outermost bar are exclusive to the respective set, while the items of the innermost bar are shared with multiple other sets. This is analogous to the magnet metaphor of Yi et al. [173] with set labels acting as magnets on the radial dimension.

Bars representing the same degree k in different histograms $\{H_j\}$ are located at the same radial position in their regions. This makes it easier to identify and interact with these bars than in Set'o'grams, where sections of the same degree are located at different heights. Furthermore, gaps in the distribution can be more easily identified, since the bars do not need to be stacked like the sections in Set'o'grams.

The bars are by default centered in their regions to avoid artificial asymmetry across the histograms and to make comparing their shapes easier. Moreover, the symmetry facilitates perceiving the histograms as figures or objects in their regions following Gestalt laws [165]. This emphasizes that these objects represent elements contained in the respective sets. A similar layout was used for augmenting histograms over the axes of parallel coordinate plots [61]. However, the lack of a baseline, the radial arrangement, and the 3D visual cues (Sect. A.3.1) impact the accuracy of comparing the length of individual bars and of estimating selected fractions of these bars (figure 1c). Therefore, Radial Sets offer an overview visualization, with precise comparisons needed to be performed on demand as we discuss in Sect. A.5.

The histogram scales can be either uniform or assigned individually to fit the histograms in regions of equal area. Uniform scaling is useful for comparing the bars of different histograms in length. Nonuniform scaling is useful for comparing the shapes of the histograms especially when the sets exhibit a large variance in size. In the latter case, the different scales can be indicated

via rectangles along the h_{jk} axes (figure A.6) scaled differently in each region to depict the same number of elements, as suggested by Cleveland [29, p. 90].

Representing the elements in each set as a histogram of their degrees gives an idea of how much overlap this set has with how many sets. This solves the tasks **T1** and **T2**. However, histograms do not tell with which sets these overlaps are. As we show in Sect. C.2.3, all 2^m possible overlaps can be analyzed on demand via interaction with the histograms. But to gain an overview of individual overlaps, additional visual elements are needed as we show in the next section.

Visualizing the overlaps

An overlap $O_{\{j_1,\dots,j_k\}} = \bigcap_{l=1}^{l=k} S_{j_l}$ is the intersection between k specific sets $\{S_{j_1}, \dots, S_{j_k}\}$ in the set system. By k we denote the degree of the overlap. Each element e in this overlap is of $degree(e) \geq k$. Hence, this overlap contains overlaps of higher degree $O_{J \supset \{j_1,\dots,j_k\}}$, and can intersect with other overlaps of degree k. The elements exclusive to an overlap $O_{\{j_1,\dots,j_k\}}$ are:

$$EO_{\{j_1,\dots,j_k\}} = \{e \in O_{\{j_1,\dots,j_k\}} : degree(e) = k\} \tag{A.5}$$

Radial Sets map overlaps to frequency-based representations of proportional size. These representations can either depict the absolute sizes of the overlaps or their normalized sizes.

$$nsize(O_{\{j_1,\dots,j_k\}}) = \frac{|O_{\{j_1,\dots,j_k\}}|}{|\bigcup_{l=1}^{l=k} S_{j_l}|} \tag{A.6}$$

Normalization makes it easier to compare overlaps between sets of different sizes by emphasizing the proportions of the respective sets they represent, as illustrated in figure A.5. Eq. A.6 computes the normalized size of an overlap by considering only the sets involved in this overlap. Disproportionality measures offer another possibility to compare two overlaps, taking into account all elements E in the set system. The disproportionality of an overlap is the deviation between the actual and expected probabilities of an element $e \in E$ to lie in this overlap:

$$disproportionality(O_{\{j_1,\dots,j_k\}}) = \frac{|O_{\{j_1,\dots,j_k\}}|}{n} - \prod_{l=1}^{k} \frac{|S_{j_l}|}{n} \tag{A.7}$$

The expected probabilities are computed by assuming marginal independence of the sets. The resulting residuals can take either positive or negative values, and can be conveyed by coloring the overlaps using a diverging color scale. Other residuals are also possible to eliminate a possible bias in Eq. A.7, caused by the sets being of different sizes [4].

To simplify overlap analysis, we restrict the visualization by default to overlaps of a certain degree k selected by the user. This is in accordance with task **T3**, where users ask questions like "which three sets exhibit disproportionally large overlap?". Moreover, this simplifies the visualization by reducing the number of visual elements needed to depict the overlaps and by making these element to have the same semantics and similar shapes. The number of possible overlaps of degree k is equal to $\binom{m}{k}$, the number of possible combinations of k objects from a set of m objects. This number can be relatively large for values of k larger than 2. Therefore, Radial Sets adopt different strategies for depicting overlaps, depending on their degrees and actual count.

Visualizing overlaps of degree $= 2$ as arcs Radial Sets visualize overlaps between pairs of sets (task **T4**) as arcs between their regions. The thickness of an arc encodes the absolute or the normalized size of the overlap (figure A.6). To alleviate clutter that results from arc crossings, the regions are ordered so that thicker arcs are kept as short as possible. For this purpose, we use a greedy algorithm that iteratively concatenates chains of regions, starting from the individual regions. At each iteration, the algorithm selects the next thickest arc between two regions and concatenates the two chains that contain these regions in one chain, optimizing on the arc length:

Algorithm 1 Compute regions' order to shorten thick arcs

for all j in $1 \dots m$ **do**
  $chain[j] \leftarrow \{j\}$ as list
end for
$overlaps \leftarrow \{O_{\{j_1,j_2\}} : 1 \leq j_1 < j_2 \leq m\}$ as list
Sort $overlaps$ in descending order of $|O_{\{j1,j2\}}|$ or $nsize(O_{\{j1,j2\}})$
for all $O_{\{j_1,j_2\}}$ in $overlaps$ **do**
  if $chain[j_1] \neq chain[j_2]$ **then**
    $c[1] \leftarrow concatenate(chain[j_1], chain[j_2])$
    $c[2] \leftarrow concatenate(chain[j_1], reverse(chain[j_2]))$
    $c[3] \leftarrow concatenate(chain[j_2], chain[j_1])$
    $c[4] \leftarrow concatenate(chain[j_2], reverse(chain[j_1]))$
    {concatenate according to the shortest arc $\widehat{j_1 j_2}$}
    $index \leftarrow \text{argmin}_i\{\widehat{j_1 j_2}$ computed in $c[i] : 1 \leq i \leq 4\}$
    $chain[j_1] \leftarrow c[index]$
    $chain[j_2] \leftarrow c[index]$
    if $|c[index]| = m$ **then** {all regions are in one chain}
      return c[index]
    end if
  end if
end for

The ordering problem resembles the seriation problem [26,92] in reorderable matrices (Sect. A.2.3). The computed order not only alleviates clutter, but also reveals clusters of sets having high overlap with each other. To analyze these overlaps more explicitly, links of higher degree are needed instead of the arcs as we explain next.

Visualizing overlaps of degree ≥ 3 as hyperedges To visualize the overlap between $k \geq 3$ sets (task **T3**), Radial Sets create a bubble of proportional size in the inner area. The bubble is connected with the respective regions via elongated arrow heads (figure 1c). The bubble along with these heads form a hyperedge over m vertices denoting the sets. To fit multiple hyperedges in the inner area, a layout algorithm is needed to reduce bubble overlaps and edge crossings. Finding the optimal solution is an NP-complete problem [40]. Therefore, we use a greedy algorithm that employs a density map to place the bubbles. The algorithm iterates over the overlaps of degree k in descending order of their absolute or normalized sizes. For each

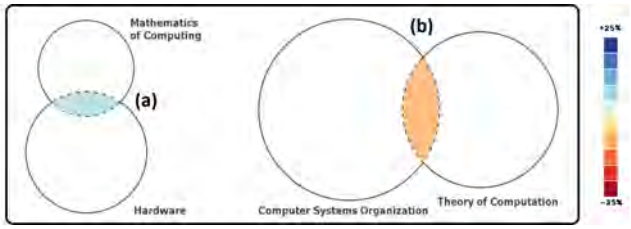


Figure A.5: Two overlaps of 2nd-degree, having different absolute sizes, but nearly equal normalized sizes (Eq. A.6). The color denotes the overlap disproportionality (Eq. A.7) using the same color scale as in figure A.3b.

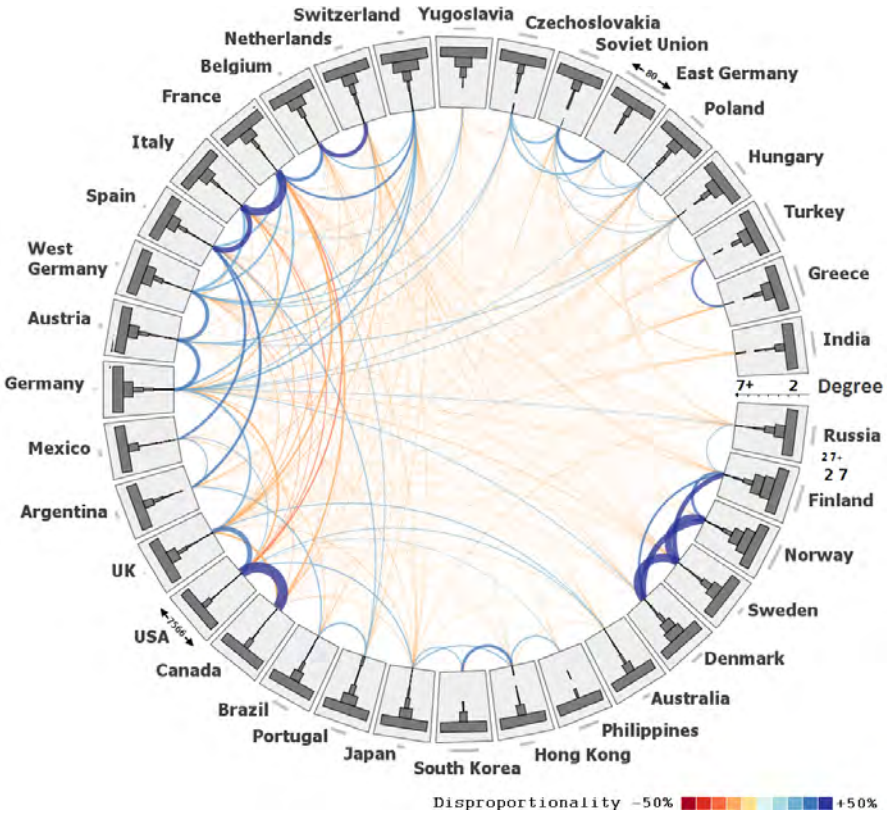


Figure A.6: Radial Sets depicting IMDb movies produced in two or more countries (including former countries). An arc between two countries represents the overlap between their movies. Its thickness and color respectively encode the normalized size (Eq. A.6) and the disproportionality (Eq. A.7) of this overlap. The different scales of the histograms are indicated as thin rectangles representing the same number of elements.

overlap it creates a hyperedge centered at a point (x, y) in the map. The point is chosen so that the overall density at the pixels the hyperedge occupies is minimized. The densities at these

pixels are increased to alleviate the overlap with hyperedges created in next iterations.

The design of the hyperedges intends to emphasize overlap sizes by mapping them to the bubble size. Bubbles are also appropriate for showing fractions of the overlaps to denote elements selected by the user (Sect. C.2.3). The edge connecting a bubble with a region is plotted with decreasing thickness to reduce clutter. The varying thickness helps to some degree in visually separating overlapping hyperedges.

Density maps have also been used to create visual links that do not occlude the visualization [144]. The algorithm described above yields interactive performance for computing the placement of 100 hyperedges with a map resolution of 200×200 pixels. The bottleneck is rather its visual scalability: hyperedges are more complex objects than arcs. This imposes a severe limit on the number of hyperedges that can be visualized with sufficient readability. Figure 1c shows about 150 overlaps of 3^{rd} degree, with the largest 10% overlaps accounting for 50% of the areas. The number and the shape complexity of the hyperedges potentially increase for overlaps of higher degree. This can rapidly increase the clutter even with a dozen sets. One way to avoid the clutter is to analyze the overlaps in a separate detail view (Sect. A.3.2). Another way is to show the links of a hyperedge only for a few number of large overlaps, or only on demand as we explain next.

Visualizing overlaps as bubbles Showing only the bubbles of the hyperedges described above results in a "bubble chartöf the overlaps. Pointing over a bubble reveals the links to the sets involved in the corresponding overlap. In case the histograms are scaled uniformly, the bubbles can be scaled using the same scaling factor. This facilitates perceiving an overlap in proportion of the involved sets. Alternatively, the bubbles can be scaled to fit in the inner area, to efficiently use this area in supporting the interaction with the bubbles and the comparison of their sizes (figure A.7).

The compactness and the uniform shape of the bubbles allow showing overlaps of multiple degrees $2 \leq k \leq b$ at once by dividing the inner area into concentric rings. Starting from the outermost, each ring k contains bubbles that represent overlaps of degree $k + 1$. A bubble can represent either all the elements in the overlap, or the elements exclusive to it (Eq. A.5). The latter case avoids the redundancy of representing the same element in multiple overlaps. The former case allows comparing absolute overlap sizes across multiple degrees to analyze, for example, the satisfaction of increasing set membership requirements. Both color and interaction allow analyzing the exclusiveness of these overlaps and the intersections they exhibit between each other, as we explain next.

Visualizing information about the elements via color

Each arc, bubble, and histogram bar in Radial Sets represents a subset of the elements E whose size is encoded by its area or thickness. Further information about the elements in this subset can be communicated by coloring this area. When the user performs a select operation over the elements (Sect. C.2.3), Radial Sets use color to depict selected fractions in each of the abovementioned subsets. If no selection exists, the user can specify which information to encode via color.

By choosing an attribute of the elements as source of the color information, the user can gain an overview of the distribution of its values in the different subsets (figure A.7). As we show in the usage scenarios (Sect. A.4), this provides insights into how this attribute correlates with the elements' membership of different sets and overlaps (task **T6**).

Color can also be used to depict *relative information* about the subsets. As can be seen in figure A.6, color reveals the disproportionality of the overlaps. Likewise, while the length of a histogram bar encodes the absolute size h_{jk} of the corresponding subset (Eq. B.1), its color can encode the disproportionality of this subset, defined as follows:

$$disproportionality(h_{jk}) = \frac{h_{jk}}{|S_j|} - \frac{k \cdot |E_k|}{\sum_{j_2=1}^{m} |S_{j_2}|} \tag{A.8}$$

In the above equation, E_k is the set of elements of degree k:

$$E_k = \{e \in E : degree(e) = k\} \tag{A.9}$$

This disproportionality measure compares the actual histograms with the ones that would result if all histograms exhibit the same distribution [2]. This reveals, for example, which sets tend to have more (or less) exclusive elements or 2nd-degree overlaps than the other sets. The exclusiveness of an overlap (Eq. A.5) can be analyzed by coloring its visual element by the average degree of its elements. An exclusive overlap receives a color that correspond to the overlap degree. Alternatively, the exclusiveness of an overlap can be analyzed via interaction, by selecting the elements E_k as we show next.

A.3.2 The Interactive Exploration Environment

The main user interface of Radial Sets comprises coordinated and multiple views that show information at different levels of detail. The *Radial Sets view* is the central part of the interface. The additional views show both summary and detailed information about the sets, the elements, and the overlaps. Together, these views enable formulating highly-expressive and visually-guided queries on the elements iteratively, and analyzing the query results in detail as we show next.

Summary views

Two views show summary information about the set system:

The *sets bar chart* depicts the set sizes $\{|S_{1 \le j \le m}|\}$ in descending order, along with the selected fractions of these sets (figure 1a). Since the sets can overlap, the bars do not sum up to the number of elements n, but to the number of their set memberships $\sum_{j=1}^{m} |S_j|$.

The *degree histogram* D (figure 1b) depicts a breakdown $\{|E_{0 \le k \le d}|\}$ of the set elements by their degrees (Eqs. A.1, A.9). The histogram bins sum up to the number of elements $n = \sum_{k=0}^{d} |E_k|$, with E_0 containing elements that belong to none of the sets of the set system. A sub histogram $D_{selected}$ depicts selected elements by their degrees.

[2] See the supplemental materials for more explanation of this measure.

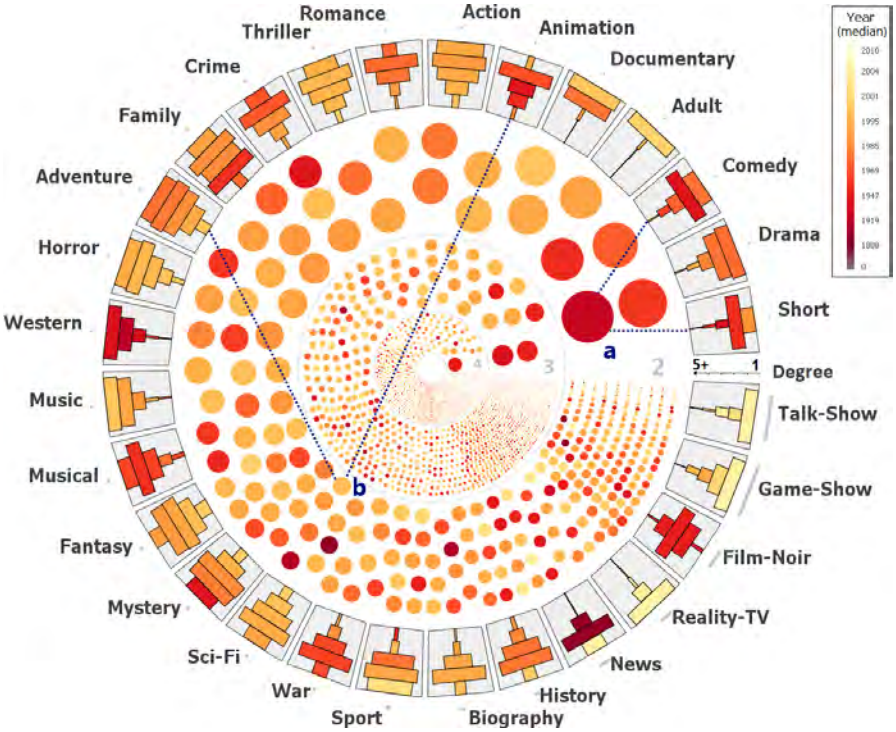


Figure A.7: Radial Sets depicting IMDb movies according to their genres. The bubbles encode the overlaps of degrees 2, 3, and 4 between the genres and are scaled to fit in the inner area. The area of a bubble encodes the normalized size of the overlap (Eq. A.6). The color represents the median release date for the movies aggregated both in the bubbles and in the histograms. The sets involved in an overlap can be inferred by hovering over the respective bubble (a, b).

Summary views are also essential to define which sets to depict in the Radial Sets view (show/hide) and which elements to incorporate in the computations (include/exclude). Furthermore, they are vital for gaining an overview on the elements under selection as well as for defining or refining the selection. Finally, both views are very useful for understanding the metaphor of Radial Sets as we explain next.

Radial Sets view

The Radial Sets view (figure 1c) can be thought of as a cross representation of both summary views: For each set S_j represented by a bar in the sets bar chart, Radial Sets show the breakdown of its elements by degree as a histogram H_j in the set's region (Eq. B.1). When the selection is equal to S_j, the sub histogram $D_{selected}$ in the degree histogram is equal to H_j, assuming no aggregation of degrees, i.e. $b = d$ (Eq. A.2).

The visual design of Radial Sets aims to provide an overview of a set system, emphasizing how the sets overlap and how the elements are distributed in them. More details about the elements and the overlaps can be obtained on demand either via tooltips or in the detail views.

Hovering the mouse pointer over a visual element in Radial Sets shows a tooltip with more information about the elements in the respective subset (figure A.8). This comprises a short description of the subset, the absolute and relative sizes of the subset and of the selected fraction in it, and further statistics such as disproportionality or aggregated attribute values. More details about the individual elements in the subsets can be obtained using brushing and linking (Sect. A.3.2).

In addition, the Radial Sets view supports direct manipulation to merge the sets or change their order using drag and drop. Merging two sets replaces them by their union and updates the visualization accordingly. The order of the sets can also be configured from the menu bar in the top of the view. The commands in this bar allow specifying color mappings (Sect. A.3.1), histogram scaling, and overlaps' degree and sizes (absolute / relative). The selection commands allow manipulating the selected elements as we explain next.

Brushing the elements for details on demand

The Radial Sets view along with the summary views expose several subsets of the elements E in the set system. Brushing these subsets enables defining a selection over E. This selection can be specified iteratively using set operations to represent a variety of combinations of these subsets. This allows a highly expressive selection of elements by their set memberships and degrees. Furthermore, the selected fractions depicted in Radial Sets and in the summary views are updated during the iterative selection. This gives an immediate feedback to the user on how the selected elements belong to the different sets and overlaps, and offers guidance on how to refine this selection[3].

Brushing the elements in a set region can be performed either by clicking on the individual bars or by defining a range over the degree axis using mouse dragging. Similar interactions are possible in the summary views and with the overlaps. If no keyboard modifier is active during the brushing operation, the selection is set to the newly brushed elements. Specific keyboard modifier can be used to specify if the brushed elements should be added to (set union), intersected with, or subtracted from the existing selection. In addition to defining the selection based on set memberships, the elements can be selected based on their attribute values. Radial Sets supports this both via textual search in the attribute values (figure 1f), or via coordinated views that enable brushing elements having certain attribute values.

The selection view shows detailed information about the selected element (figure 1d). The top of this view shows a formula that details how the selection was specified. The formula text is composed using the common set-theory notation, with extensions to express further conditions on the elements' degrees and attribute values. The body of the selection views is a tabular list of the elements in the selection, showing their attribute values. The list can be sorted by one of these attributes. These attributes can also be analyzed in detail via additional views (figure 1g). Clicking on an element in the tabular list highlights this element and shows its set memberships both graphically and in text. The text is shown at the bottom of the selection view as a comma-separated list of these set memberships. Additionally, these memberships are

[3]The supplementary video demonstrates the interactive selection of elements in Radial Sets in detail.

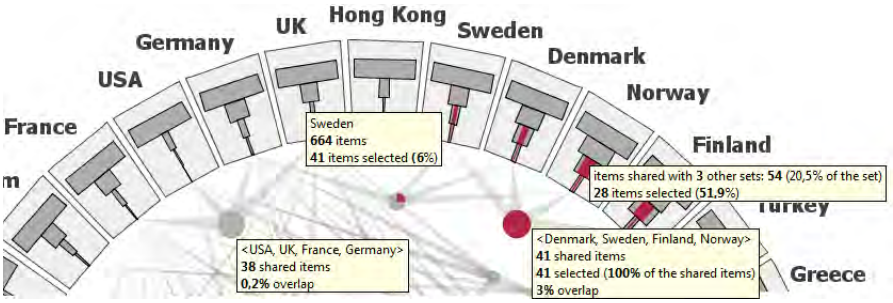


Figure A.8: Tooltips showing various information about the subsets represented by the regions, the bars and the links in Radial Sets.

indicated graphically as a star graph over the Radial Sets view. This graph shows in which region and in which bars in these regions the highlighted element is present.

Besides gaining details into specific elements, interactive selection is also useful for filtering and manipulating the data. it can be used to hide or exclude certain elements from the analysis based on their attributes, degrees, and set memberships. This is useful for dealing with real-world datasets that often exhibit highly skewed distributions of set sizes (few sets comprise the majority of the elements) or of element degrees (most elements are exclusive in their sets). Filtering out such elements reveals finer details about the rest of the data.

The expressive power of the interactive selection possibilities and the immediate feedback on selected fractions in Radial Sets, enable an elaborate analysis of the set memberships and the attribute values of certain elements in the set system (task **T7**). These possibilities constitute a visual query language for set-typed data. This language covers all possible 2^m overlaps between the sets, and goes beyond by allowing the selection of exclusive parts of these overlaps, parts having specific degrees, or parts containing certain values for selected attributes. Furthermore, the memberships of selected elements in different overlaps can be analyzed in detail as we explain next.

Overlap analysis view

The arcs and bubbles in Radial Sets give a compact overview of existing overlaps and the sets involved in them. They are also suited to revealing overlap patterns such as clusters of highly overlapping sets and to quickly select a specific overlap. To analyze and compare the overlaps in more detail, Radial Sets employ a coordinated view that shows these overlaps in tabular lists (figure 1e). Each list L_k in this *overlap analysis view* contains the overlaps of a specific degree $k \geq 2$. An additional list L_1 contains the sets like in the summary sets bar chart (Sect. A.3.2), along with further statistics about the sets. For each overlap $O_{\{j_1,\ldots,j_k\}}$, the list L_k textually shows the sets $\{S_{j_1}, \ldots, S_{j_k}\}$ involved in this overlap, separated by commas and ordered by their order in L_1. Additionally, L_k can show the absolute and normalize sizes of the overlap, the fractions of selected elements in it, a summary value of the color attribute in the whole overlap and in the selected portion, and the disproportionality of the overlap and of its selected portion. These statistics can be shown either textually or graphically using color and/or bar charts. The

overlaps list L_k can be sorted according to these statistics. This enables a detailed analysis of the overlaps in the lists and quickly finding large or overrepresented overlaps at different degrees, without having space limitations or clutter issues.

The overlap analysis view is interactively updated when the selection changes. Also, the Radial Sets view is updated when an overlap in one of the lists $L_{k\geq 2}$ is clicked: In case the view already includes a visual element for this overlap, it becomes highlighted. Otherwise, a new visual element is overlaid in the Radial Sets view to indicate involved sets and the size of the overlap in proportion to them.

A.4 Usage Scenarios

To demonstrate Radial Sets, we report insights we gained in two real-world set-typed datasets using some of the features described in Sect. A.3. The datasets are of different scales and skewness, and deal resp. with multi-label classifications and with multi-valued attributes.

A.4.1 ACM Paper Classification

The ACM digital library comprises computer science papers tagged with multiple index terms from the ACM classification system [1]. We define a set system over a collection of more than $50,000$ ACM papers extracted by Santos and Rodrigues in 2008 [127] . The sets of this system are the top-level index terms ($A.$ to $K.$), also called classes. Figure A.3b depicts the Radial Sets of these index terms. Each histogram bar is colored by the median publication date of the papers it represents. The arcs depict the overlaps between the index terms, with thickness and color representing the normalized size (Eq. A.6) and disproportionality (Eq. A.7) respectively. From the histograms it can be easily seen that the index terms vary in their exclusiveness: few computer-science papers are exclusive to class G (Mathematics of Computing); while 92.2% of the papers in this class have other index terms. On the contrary, 42% of "Hardware" papers did not have other terms assigned.

It is also noticeable that the index terms vary in the recency of their papers, indicated by the median publication date. The median date varies between 1994 (classes F and G) and 2001 (classes C and E). Also, papers that belong to one class tend to be more recent than papers that belong to multiple classes, with medians at 2003 and 1997 respectively. This variance can be easily inferred by coloring the bars in the summary charts (Sect. A.3.2) with the median dates. However, by examining the Radial Sets view, finer details about this variance can be observed, compared to the summary views. For example, contrary to the global trend, papers exclusive to class G have a median date of 1984, which is significantly older than the class median 1994. On the other hand, while class J has also a relatively old median date of 1995, the small fraction of papers exclusive to it have a very recent median date of 2005. A similar contrast between exclusive and shared papers is noticeable in class C. To verify the above observations, we plot the distribution of publication date in each of these paper classes as histograms, along with sub-histograms that represent the papers exclusive to them (figure A.9). This confirms the recency trend of class C with exclusive papers in this class being an increasing trend, constituting 67% of the papers in 2007 (up from 10% in 2000). A similar observation holds for papers

exclusive to class J: they started to appear in 2002, and made up 40% of "Computer Application" papers in 2007. To get more details about these papers, we select them in the Radial Sets view and examine the venues they were published in using the detail view (Sect. A.3.2). Most of them were published in conference series that started in the past decade on topics like "mobile computing", "genetic and evolutionary computation", "electronic governance", "future play", and "advances in computer entertainment technology" to mention a few. The long tradition of class G is observable, with papers exclusive to it being an old trend that disappears in the 1990s and reappears in the past decade. To investigate this trend, we select the G-exclusive papers whose publication dates are newer than 2000 and observe their venues. While some of these venues are recent like "Symbolic-Numeric Computation", the majority of them are established yearly conferences that were started in the 1980s or earlier on topics like "symbolic and algebraic computation", "theory of computing", "computational geometry", "parallelism in algorithms and architectures" and "supercomputing". By searching for all papers of these conferences in the dataset and examining their publication dates, we consistently found full or large gaps in the 1990s. This explains the gap we observed for the G-exclusive papers (figure A.9) and reveals a sampling bias in the dataset.

The insights gained so far are focused on set-membership tasks (**T1** and **T2**) and attribute-analysis tasks (**T6** and **T7**). To explicitly analyze set overlaps (tasks **T3**, **T4** and **T5**), we observe the arcs in figure A.3 and the hyperedges in figure 1c. From the arcs we immediately notice a significant overlap between "Mathematics of Computing" and "Theory of Computing". This overlap constitutes 15.5% of the union of these classes; up from 5% expected overlap in case of statistical independence. Many other disproportionally-high overlaps are noticeable such as "Information Systems" ∩ "Computer Methodologies" and "Hardware" ∩ "Computer Systems Organization". On the other hand, there are classes that exhibit only a small overlap such as "Hardware" ∩ "Information Systems". By examining the 207 papers in this overlap in the detail views, we observe that many of them were published in conferences on "Design Automation" (40 papers), "Human Factors in Computing" (24), and "Management of Data" (19).

Hyperedges with large bubbles in figure 1c indicate significant overlap between three classes, such as $D \cap H \cap K$ and $F \cap G \cap H$. In this figure, papers having "Human Factor" in their general terms are selected, comprising about 19.6% of the dataset. The bubbles are colored to indicate selected fractions in the overlaps. Certain overlaps have disproportionally-large selected fractions. For example, 66% of papers on "Computing Milieux", "Computing Methodologies", and "Information Systems" address issues of "Human Factors". This ratio is higher in the overlap than in its individual classes, as can be observed in the summary view (figure 1a). These papers were published in conferences like ACM CHI (48 papers), SIGACCESS (40), SIGCSE (26) SIGGRAPH (16), and IUI (9).

A.4.2 IMDb Movies

Information about movies comprises several multi-valued attributes such as genres, production countries, and languages. To illustrate the insights gained by Radial Sets in such attributes, we consider two set systems that can be defined over a 2010 snapshot of the IMDb database [2] comprising over 525,000 movies.

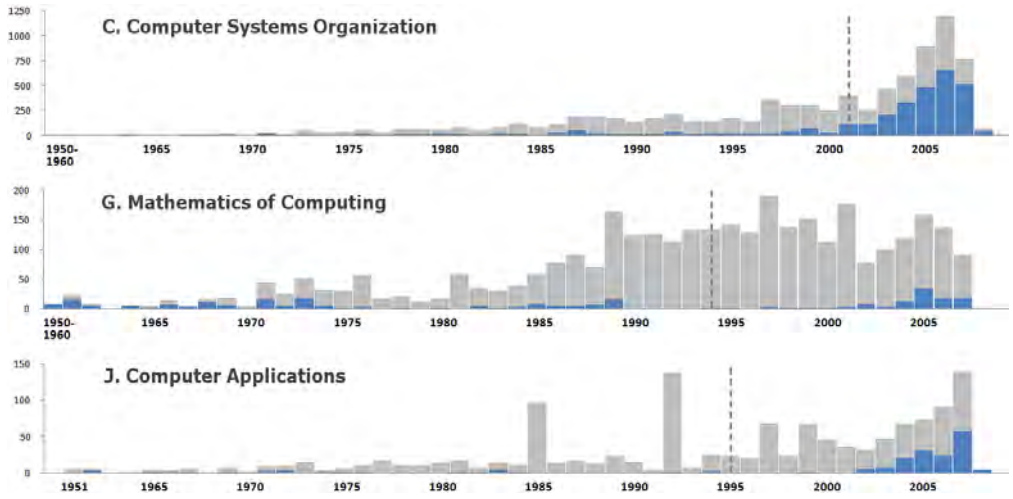


Figure A.9: The number of papers over time for three classes in the ACM digital library. Exclusive papers in each class are highlighted in blue. The dashed lines indicate the median publication date in each class.

The sets of the first system are the top 35 production countries of the movies. The sets exhibit a large skewness in their sizes with the US being involved in 38% of the movies, followed by the UK (7.7%). The smallest sets are East Germany and Russia, each involved in about 0.4% of the movies. Another large skewness exist in the distribution of the element degrees: 96% of the movies were produced in one country. These elements do not contribute to any overlaps, and hence are less important for analyzing co-production patterns between the countries. Including them obscures finer information about the overlaps. Similarly, very few movies (0.03%) were produced in five or more countries, with only one movie having the largest element degree of 13. Therefore, we group elements of degree ≤ 5 to increase the resolution of the histogram bars. Depicting absolute values in the histograms will assign the majority of the available space to the few top-5 countries and obscure the rest of the data. Therefore, we assign the regions equal areas to enable relative comparison of the distributions in these histograms (figure A.6). This reveals a variety of patterns in the data: pairs of countries that produced *relatively* more joint movies than other pairs become visible (**T4**). Such countries often have a common language or a common border. The ordering algorithm reveals groups of countries that exhibit high mutual overlaps, most noticeably the Scandinavian countries. By checking the 4^{th}-degree overlaps in the overlap-analysis view, we immediately notice that 41 movies were produced jointly by all of Denmark, Finland, Norway, and Sweden, making this the largest overlap of 4^{th} degree (**T3**). Figure A.8 shows the absolute sizes of these overlaps graphically using hyperedges. The 2^{nd}-largest overlap is between USA, UK, France and Germany, the four largest sets comprising 56.5% of all movies. This points to a very significant disproportionality of the Scandinavian overlap, given the small sizes of the involved sets (summing up to 3.5% of all movies).

The sets of the second systems are the 28 IMDb movie genres. Figure A.7 depicts the Radial Sets of the genres set system. The bubbles in the different rings represent normalized overlaps of

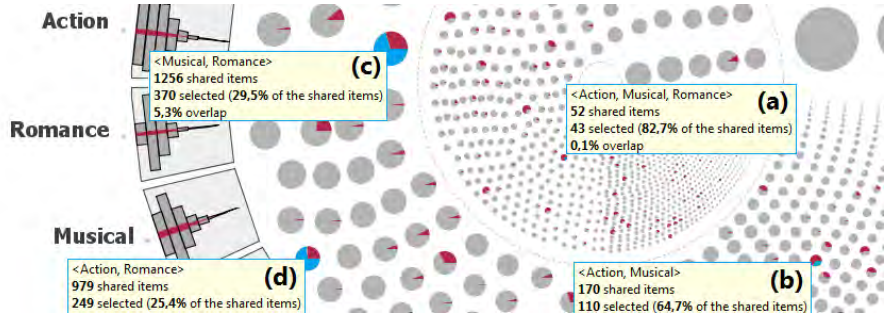


Figure A.10: Selected genre overlaps that exhibit disproportionately high presence of Indian movies (highlighted in red).

degree 2, 3, and 4. We also employ relative analysis both for the histograms and for the bubbles due to the high skewness between the set sizes. We easily notice that the genres vary in their exclusiveness (**T1** and **T2**): 94.1% of Animation movies had other genres, whereas 93.1% of Adult movies were exclusive to this genre. The elements are colored by the median release date of the movies they represent. This reveals a significant variance in the recency of the genres and their combinations (**T6**). For example, movies exclusive to Mystery were predominantly old (median date 1944), whereas Mystery movies that have other genres are more recent (median 1988). The opposite holds for genre News. This is revealed by contrasting the first bar with the other bars in the regions.

The combination Comedy ∩ Short contains mainly the older movies (median 1926) from both genres (figure A.7a) that have individually more recent median dates (1967 and 1966). In contrast, Animation ∩ Adventure contains mainly the newer movies (median 1997) from both genres (figure A.7b) that have older median dates (1966 and 1986).

The two set systems (countries and genres) can be analyzed against each other to find disproportionalities in the overlaps (**T7**). For example, while Indian movies comprise 2% of the dataset, selecting them in the Radial Sets of Genres reveals higher percentages in specific overlaps (Fig A.10). In particular, these movies comprise 83% of Musical ∩ Action ∩ Romance (figure A.10a), and 65% of Musical ∩ Action (figure A.10b). The other two pairs of these three genres (figure A.10c-d) exhibit less percentages. These findings can be analyzed and compared against each other in more details in the overlap analysis view.

A.5 Discussion

Radial Sets build upon and extend several ideas from state-of-the-art techniques to enable advanced visual analysis of large overlapping sets. Our technique extends the frequency-based aggregation of Set'o'grams [48], which accounts for high scalability in the number of elements of the set system. Also, it uses separate visual elements for the sets and for the overlaps, similar to the untangled Euler diagrams [63]. The hyperedges between radially-arranged regions are inspired from the free nodes in Anchored Maps [106]. The radial layout is adopted from Contingency Wheel++ [4] which was designed to visualize skewed contingency tables having few

columns but a large number of rows. These tables have a similar structure and dimensionality as the elements-set membership matrix. Nevertheless, Radial Sets use different aggregation for the elements in the histograms and in the overlaps, and introduce additional visual elements to address the characteristics of set data and support the tasks specific to them.

The visual design of Radial Sets is a compromise between information richness and effectiveness. For example, an $m \times m$ heatmap can be more effective at showing the 2nd-degree overlaps than crossing arcs with a limited range of varying thicknesses. Also, standard bar charts are more precise at showing the elements by degree in each set than non-aligned bars depicted in radially-arranged parallelograms. Finally, color is sub-optimal for showing the values of an attribute in the elements aggregated in a bar or in a bubble. Nevertheless, depicting all this information together enables gaining a high-level overview of the distributions of the elements, the overlaps, and the attributes, in relation to each other. Using separate visualizations such as an overlap matrix, element histograms and attribute histograms makes it harder to visually link between related elements. Our interaction techniques allow certain elements in Radial Sets to be investigated at greater detail on demand using simpler and more precise visualizations. Hence, Radial Sets serve as a starting point of the analysis and as a means to detect extreme differences and to quickly formulate queries to select these elements. However, in an informal pilot feedback session with 10 engineers from different disciplines, three subjects reported that Radial Sets of movies are showing too much information at once. One of them recommended showing the arcs for one selected set only. Nevertheless, the subjects were able to interpret the visual metaphor correctly and use interaction to perform set operations on the elements and to answer questions on the relations between the sets.

The visual complexity of Radial Sets imposes a limit on the number of sets it can depict. For example, using the 2$^{\text{nd}}$-level classes of the ACM classification in Sect. A.4.1 results in 89 sets, each receiving 4 degrees of the angular resolution on average. A higher angular resolution is needed to ensure a sufficient readability of the histograms, which limits the number of sets that can be visualized at once to about $m \leq 30$. On the other hand, Radial Sets can handle a large number of elements, at the order of 1 million, thanks to the frequency-based aggregations and to the relative analysis possibilities. For example, figure A.7 depicts information about 525,000 movies using non-uniform scaling for the histograms and normalized sizes for the bubbles.

Another limitation is the number of hyperedges that can be visualized at once being ≤ 100 (assuming a normal distribution of overlap sizes), which is only 2% of all possible 3$^{\text{rd}}$-degree overlaps between 30 sets. The remaining possibilities in these cases are to show the bubbles only or to analyze the overlaps separately in the detail view.

Finally, using separate visual representations for the overlaps and for the sets hinders the depiction of containment relations between the sets. Such relations are pre-attentive in Euler diagrams [82], even in a composite form such as $S_1 \subset S_2 \cup S_3$ or $O_{\{1,2\}} \subset S3$. Also, both the arcs and the bubbles show the absolute or normalized size of an overlap, without indicating the different fractions it constitutes in the involved sets. This information needs to be investigated on demand by selecting this overlap and checking these fractions individually.

Future Work One way to compensate for the visual limitations of Radial Sets and their low sensitivity to small differences between attribute values or overlap sizes is to employ complementary computational methods. These methods can pre-compute significant disproportionalities in the overlaps and in attribute distributions among all elements or in selected subsets. We are investigating statistics-based and computationally-efficient measures for this purpose along with possibilities to communicate their results visually, and steer the calculation interactively. Additionally, we are considering different placements of histograms and hyperedges to address the perceptual issues of the current layout as well as alternative visual representations based on heatmaps to visualize the same information in a more scalable way in the number of sets. Finally, to confirm our informal findings on the understandability of our visual design, we are currently conducting a formal evaluation of Radial Sets that will assess how well they support the tasks intended in comparison with other alternatives.

A.6 Conclusion

Radial Sets is a novel interactive technique for the visual analysis of large overlapping sets, designed to provide insights into different kinds of overlaps between the sets. These overlaps are salient features of set-typed data and are central to relevant analysis tasks. Radial Sets builds upon selected ideas from existing techniques to support these tasks in a scalable way using several aggregation methods and a multi-level overview+detail exploration environment. In particular, our technique enables (1) gaining insights into different kinds of overlaps between the sets and into the disproportionalities they represent, (2) analyzing the element memberships of the sets and the overlaps in relation to other attributes of the elements, and (3) interactively querying the elements by their set memberships and attribute values, and analyzing how selected elements differ from the rest of the elements in their memberships of the sets and the overlaps. As the usage scenarios demonstrate, Radial Sets enable conducting elaborate analysis workflows in large set-typed data using expressive visual queries. These queries allow combining set-theoretic operations to select and analyze specific elements of interest in the data. Compared with existing visual representations, Radial Sets offer richer information in the overview but at lower precision and sensitivity to small differences. Nevertheless, using interaction and complementary views, Radial Sets reveal a variety of overlapping patterns in large overlapping sets, beyond the limits of state-of-the-art techniques.

CHAPTER B

Visual Analytics Methods for probabilistic classification Data

Appears in IEEE Transactions on Visualization and Computer Graphics, 20(12), 2014.

Abstract: Multi-class classifiers often compute scores for the classification samples describing probabilities to belong to different classes. In order to improve the performance of such classifiers, machine learning experts need to analyze classification results for a large number of labeled samples to find possible reasons for incorrect classification. Confusion matrices are widely used for this purpose. However, they provide no information about classification scores and features computed for the samples. We propose a set of integrated visual methods for analyzing the performance of probabilistic classifiers. Our methods provide insight into different aspects of the classification results for a large number of samples. One visualization emphasizes at which probabilities these samples were classified and how these probabilities correlate with classification error in terms of false positives and false negatives. Another view emphasizes the features of these samples and ranks them by their separation power between selected true and false classifications. We demonstrate the insight gained using our technique in a benchmarking classification dataset, and show how it enables improving classification performance by interactively defining and evaluating post-classification rules.

Keywords:Probabilistic classification, confusion analysis, feature evaluation and selection, visual inspection.

B.1 Introduction

The performance of classifiers in terms of correct classification is a major factor in determining their applicability for a given problem. Significant advances in machine learning have led to the development of a variety of classifiers and to an improved understanding of their properties. Designing a classification algorithm for a given problem is usually an iterative process that

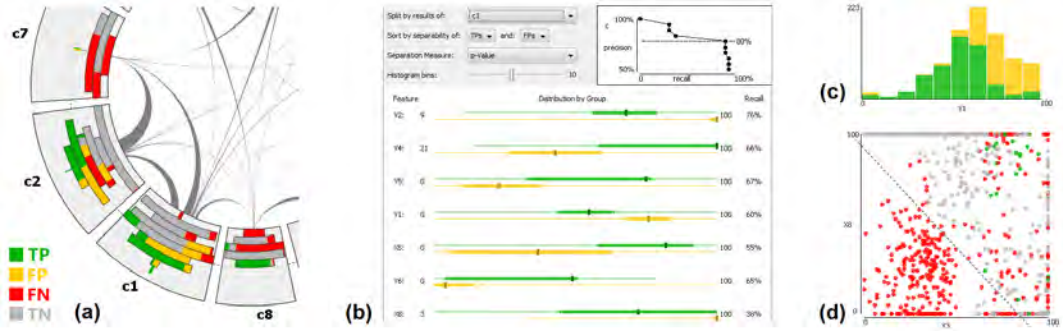


Figure B.1: Our visual analysis tools: (a) the *confusion wheel* shows sample-class probabilities as histograms colored by classification results, (b) the *feature analysis view* depicts feature distributions among selected samples, separated by their results, and ranked by a separation measure, (c, d) histograms and scatterplots reveal the separability of selected true and false classified samples by one or two features.

involves several decisions and choices. These include choosing an appropriate classifier, parameter tuning of this classifier, feature selection, and possibly introducing specific extensions to the classifier in order to handle special cases or to incorporate domain knowledge. This process aims to optimize the performance of the classifier according to some measures such as error rate, cost, or risk. For each of the above-mentioned stages in the design process, machine-learning experts need to understand the data involved in order to make choices that increase performance. Providing tools that assist these experts in analyzing the data in relation to the classification performance enables valuable guidance for the design process [88, 163].

Visualization has played an important role in understanding and comparing classification algorithms and in improving their design (Sect. B.2). In case of multi-class classifiers, the performance is usually reported by means of a confusion matrix that records for each class how many times its samples were confused for each other class (Fig. B.2a). Compared with overall performance measures, these matrices provide more details about the results and help in introducing appropriate adjustments to the classifier. Besides predicting the class for a given input sample, many multi-class classification algorithms compute likelihood scores for a sample being of each of the classes. Analyzing how these scores correlate with the classification error and data features is important to understand the behavior of such classifiers. Confusion matrices discard this information as they incorporate final classifier decisions only. This paper presents visual methods for analyzing classification results of a multi-class probabilistic classifier for a large number of labeled samples. After motivating this problems and identifying related tasks (Sect. B.3- B.4), we describe how our methods allow analyzing classification results in context of class probabilities (scores) and data features. Our main contributions are:

- Involving class probabilities in the analysis of classification data by explicitly representing them as colored histograms.
- Intertwined automated and visual methods to analyze data features in relation with classification results and probabilities, and to rank them by their separation of true and false

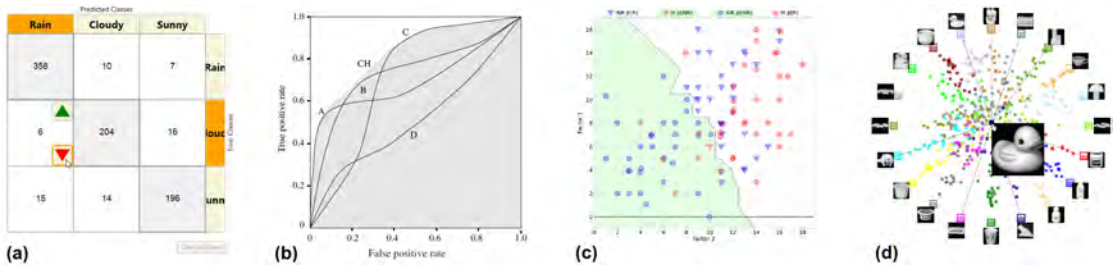


Figure B.2: Techniques for visualizing classification results: (a) an interactive confusion matrix [77], (b) ROC curves comparing five classifiers [43], (c) binary classification boundaries projected on two factors [103], (d) Class Radial Visualization [134].

classification.

- An interactive exploration environment to analyze different aspects of probabilistic classification data.

Sect. B.5 presents usage scenarios to demonstrate the applicability of our methods in analyzing and improving classification results. In Sect. B.6 we discuss the advantages and shortcomings of our approach compared to previous work, and report expert feedback as well as practitioners experience in analyzing their classification data.

B.2 Related Work

A variety of approaches and tools have been proposed to improve classification performance using visualization. They can be categorized into techniques that engage the user actively in building the classifier, and those that focus on retrospective analysis of the performance.

B.2.1 Building Classifiers Interactively

Ware et al. [163] argue that classifiers built by users can compete with automated techniques as the users can incorporate their domain knowledge in the classifier design. Several techniques have been proposed for interactively constructing specific classifiers such as ones based on linear discriminant analysis (LDA) [27] or decision trees [11, 150, 158]. Also, similar ideas were proposed for specific aspects of classifier design such as distance measures [10, 22] and feature selection [21, 98, 174].

Certain techniques offer visual support for active learning, an approach which enables machine learning algorithms to query the user for the desired class of an unlabeled sample [135]. This learning paradigm has been shown useful in several domains such as video analysis [67] and document retrieval [62] where the number of samples is very large, prohibiting a manual labeling beforehand [133].

Talbot et al. [147] presented an interactive system to support ensemble learning, an approach to combine multiple classifiers to build one that is superior to its components. Their

EnsembleMatrix technique visualizes the confusion matrices of the individual classifiers and allows combining these classifiers interactively with immediate update to the combined confusion matrix. Kapoor et al. [77] developed *ManiMatrix*, a system to refine a classifier by means of simple interactions with the confusion matrix (Fig. A.2a). Reducing the tolerance for confusion between two classes triggers a search for new classification boundaries and updates the matrix interactively if a solution is found.

B.2.2 A Posteriori Analysis

Several visualization techniques were developed to help machine-learning experts analyze classification results *a posteriori*. These techniques are not tightly integrated with the classifiers, but are designed for post-mortem analysis, which makes them potentially classifier-agnostic. We describe next three categories of these techniques according to the primary information they visualize.

Classifier performance: Receiver operating characteristic (ROC) curves [43] and their variations [39] are well-established methods for tuning, assessing, and comparing the performance of binary classifiers. A ROC curve plots the true positive rate against the false positive rate of a binary classifier for a varying discrimination threshold (Fig. A.2b). While ROC analysis can be extended to multi-class classifiers, it is still computationally exhaustive due to the large number of class combinations that need to be computed [88]. Also, visualizing high-dimensional ROC spaces is challenging, with existing techniques being able to show only partial information about the classes [59].

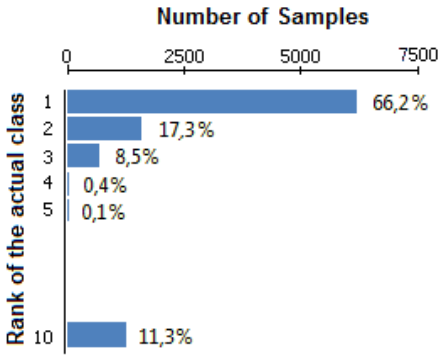
Data features: These techniques are dedicated to analyze the influence of data features on the classification results. Anand et al. [10] use a bubble chart to depict how the samples of a target class are distributed based on the values of one nominal and one numerical data features. Kienreich and Seifert [83] use feature-class matrices to show how features correlate with classes. A number of techniques visualize the decision boundaries of a binary classifier in a multi-dimensional feature space [23, 103]. Multiple scatter plots of the data features are used for this purpose (Fig. A.2c). A recent follow-up technique augments the data points with information about their distances to the decision boundary [102]. This was shown useful for steering the classification model and identifying cost-changing data elements.

Class probabilities: Dedicated techniques have been proposed to visualize class probabilities in the case of probabilistic classification. Rheingans and desJardins used a heatmap to visualize the probability of a given class for each value combination of two features [121]. They account for a larger number of features by creating a 2D projection of the feature space. Iwata et al. [71] proposed a projection of class probabilities to visualize multiple classes in the same time. Projection-based techniques can preserve interesting structures in the high-dimensional space. Nevertheless, they might potentially result in complex visualizations that require good understanding of their properties and semantics to interpret correctly. Seifert and Lex [134] proposed a simplified technique for visualizing class probabilities. It places the classes on a circle

and depicts the samples as points in this circle based on their class probabilities (Fig. A.2d). These probabilities can be shown on demand for one sample as lines of varying thicknesses. The points are colored according to their predicted classes. In our work we also employ a radial layout for the classes, but use different visual abstractions and interactions as we explain next.

B.3 Motivation of our Work

When classifying samples using a probabilistic classifier, it is possible to infer for a wrongly-classified sample whether the actual class is the 2nd, 3rd or last guess (i.e. the rank of the actual class). Fig. B.3 shows a histogram of the ranks computed by a classifier for the actual



classes of 10,992 samples. About 17.3% improvement on the classification rate is possible if the classifier would succeed on the 2nd guesses, e.g. by simple adjustments to the classifier or by using additional classification rules. On the other hand, 11.3% of the samples fail with low improvement chance with the current classifier. These samples are hard to separate from non-class samples.

The histogram of actual class ranks does not provide actionable insight beyond indicating the amount of improvement potential. More detailed visualizations are needed to guide the users on how they can improve the performance. Fig. B.3a shows a confusion matrix of the classification results described above. The size of a cell encodes the samples of its row class that are confused for its column class.

The matrix is augmented with histograms of sample probabilities in each row and column. The row histograms represent false negatives (FNs), while the column histograms represent false positives (FPs) in the respective class. As we show in the next sections, this information is vital to understand the behavior of probabilistic classifiers. However, the matrix representation does not assign visual primacy to the histograms, which limits their usefulness. Moreover, it does not include information about correctly classified samples (true positive TPs and true negatives TNs) and how their probabilities are distributed, compared to misclassified samples. Finally, the information related to one class is scattered in multiple cells and two histograms in the respective row and column. We address these issues by employing alternative visual designs dedicated for analyzing probabilistic classification data.

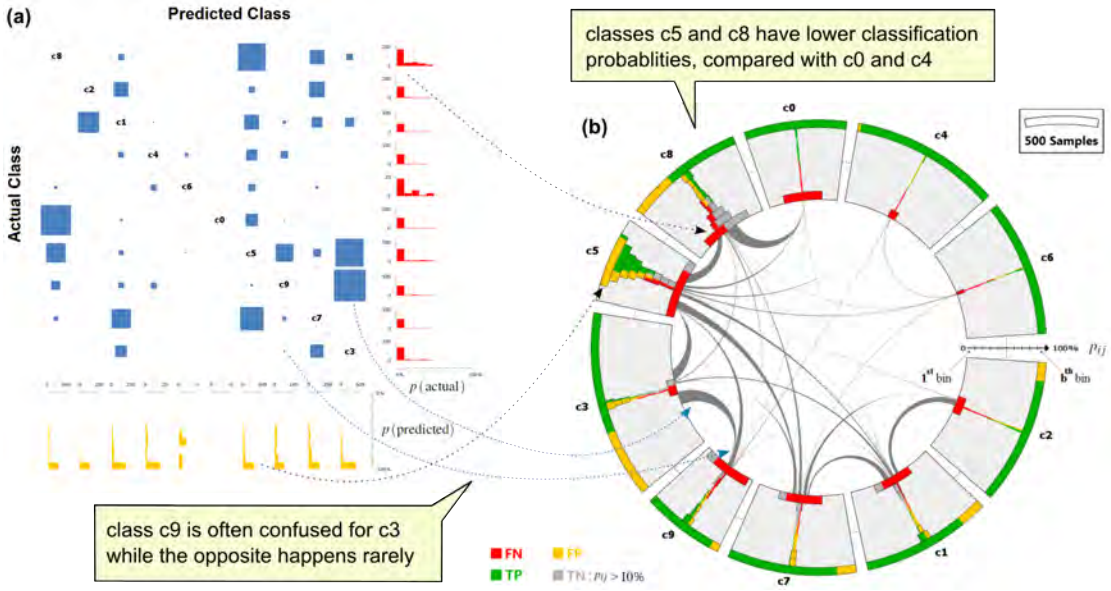


Figure B.3: Visualizing classification results of 10,992 handwritten digits [12] (a) using a confusion matrix augmented with histograms of sample probabilities in the respective rows and columns, (b) using the confusion wheel: Sectors represent digits with chords showing classification confusion between them. Histograms represent the probabilities of the samples in each class according to the color legend.

B.4 Our Visual Analysis Tools

We propose a set of visualization tools that are integrated to analyze probabilistic classification data. The data encompasses:

- A set S of n labeled samples that are classified into m classes $C = \{c_{1 \leq j \leq m}\}$.
- The actual label for each sample $l^a(s) \in C : s \in S$.
- The predicted label for each sample $l^p(s) \in C : s \in S$.
- The probability p_{ij} for each sample $s_i \in S$ to belong to class $c_j \in C$, as computed by the classifier.
- A set of l data features $f_{1 \leq k \leq l}(s_i)$ for each sample $s_i \in S$, used by the classifier to compute the class probabilities.

The above information is available when classifying unknown samples, except for the actual labels l^a. Therefore, l^a-independent observations in the data can be reproduced during actual classifications.

Our tools aim to support the following analysis tasks:

- **T1:** Analyze the overall probability distribution of the samples to belong to a class (regardless to their actual classes).

- **T2:** Compare the probability distribution of the samples according to their classification results (TPs, FPs, FNs, TNs) in a class.
- **T3:** Find out FNs / FPs that have high / low probability. These samples are easier to improve than other FNs and FPs.
- **T4:** Select samples confused between two classes, and analyze their probability distributions in these classes.
- **T5:** Select samples by their class probabilities, classification results, or data features for further analysis.
- **T6:** Find out if FPs / FNs at a certain probability range can be separated from TPs / TNs in that range by the data features.

All of these tasks involve the class probabilities, and some of them involve the data features as well. Matrix representations fall short of supporting these tasks, as they assign visual primacy to class confusions. Therefore, we propose two main visualizations that assign visual primacy to the probabilities (Sect. B.4.1) or to the data features (Sect. B.4.2). We show in the next sections how these views are suited for solving the above-listed tasks, and enable new insight in the classification results beyond the information available in the matrix representation.

We illustrate our tools based on a UCI benchmarking dataset [12] that contains 10,992 labeled samples representing pen-based handwritten digits. The samples have 16 data features that comprise the x and y coordinates of eight points sampled along the curve of each digit.

B.4.1 Confusion Wheel

To assign visual primacy to the sample-class probabilities, we create histograms to show how they are distributed in each class (task **T1**). We employ the visual layout of Contingency Wheel++ [4] which places these histograms in a ring chart whose sectors represent the classes $c_1..c_m$ (Fig. B.3b). In contrast to the matrix, this layout emphasizes the classes as primary visual objects with all information related to a class grouped in one place. This includes class probabilities and confusions with other classes as we explain next.

Visualizing sample-class probabilities as histograms

For each class c_j, the samples S are divided into four sets according to their classification results:TP_j, FP_j, TN_j, and FN_j. A histogram of the class probabilities is created for each of these sets using a uniform number of bins b, equal to n by default. The samples X_{jk} aggregated in bin k in this histogram are a subset of the corresponding set X_j, where X_j is one of the four sets mentioned above:

$$X_{jk} = \{s_i \in X_j : (k-1)/b < p_{ij} \leq k/b\} \tag{B.1}$$

A closed interval $[0, 1/b]$ is used for the first bin. The confusion wheel visualizes these histograms along the radial dimension in the respective sector, with the first bin placed next to the inner ring and the last bin b next to the outer ring. The user can select which histograms to include in the visualization (task **T2**). These histograms are stacked and centered in each sector to show the probability distribution of the respective samples. Color reveals the breakdown of

these samples according to their classification results (Fig. B.3b). All histograms have the same scale and the sectors are scaled to fit them.

By default, the confusion wheel filters out the bottommost bars of TNs having $p_{ij} \leq 10\%$. Showing these samples is of marginal interest, as they usually do not compete with the winner class. Filtering out these samples increases the resolution of the other histograms that show more important information about TPs and misclassified samples. Likewise, it is also possible to filter out the topmost bars of TPs having $p_{ij} \geq 90\%$ to further increase the resolution.

A reference circle indicates the 50% probability level in each sector. All negatives are located below this line. It is also possible for positives to lie below this line: this happens, for example, when the highest three class probabilities for a sample are nearly equal.

The colored histograms give more information about the classifier performance than confusion matrices. For example, it is evident in Fig. B.3b that classes c_4 and c_6 have the clearest discrimination, with the majority of positive samples ($\geq 98\%$) in these classes being predicted with high probabilities ($\geq 90\%$). The opposite holds for c_5, where only 48% of its positive samples predicted with ($\geq 90\%$) probabilities. Only 25% of these samples were classified correctly. The percentage information can be obtained interactively in a tooltip.A naïve Bayesian classifier is used to classify the data.

Fig. B.4 shows three classes from the data depicted in Fig. B.3b. It includes only misclassified samples (FPs and FNs) depicted at a higher resolution by filtering out TPs and TNs (task **T3**). The samples are colored according to their actual classes. For this purpose, a unique color is assigned to each class from an appropriate qualitative color scale. As a result, the FNs in each class are colored by the class color. The FPs are colored by their actual classes, showing which other classes were confused for this class, and at which probability. This reveals that samples confused for c_3 are mostly of classes c_5 and c_9. Also, there is mutual confusion between c_5 and c_8 in the probability range $]0.2, 0.8]$. A misclassified sample s is double coded in Fig. B.4 since it counts as a FP in $l^p(s)$ and as a FN in $l^a(s)$. This is indicated by the chords that represent class confusions are we explain next.

Visualizing class confusions as chords

The confusion wheel depicts class confusions as chords between the sectors (task **T4**). A chord between two classes is depicted with a varying thickness: the thickness at sector j_1 is proportional to $M_{j_2 j_1}$, the number of elements of class c_{j_2} confused for c_{j_1}. Likewise, the thickness at sector j_2 is proportional to $M_{j_1 j_2}$. Hence, following the chords outgoing from a sector reveals for which other classes its false negatives are confused. Alternatively, the chord can be split into adjacent ones that show the confusion in each direction individually. The sectors are ordered so that thicker chords are made shorter, using an $O(m^2)$ greedy algorithm [5]. This reduces the visual ink and the clutter caused by chord crossings, resulting in a clearer visualization. Also, this reveals groups of classes that have more confusion among each other than with the other classes. For example, it is evident in Fig. B.3b that the digits 1, 2 and 7 are often confused with each other, as their shapes are similar to some degree.

Compared with the chords, a matrix representation is more accurate at showing class confusions and their distribution in the matrix. Nevertheless, a chord is better at showing the mutual

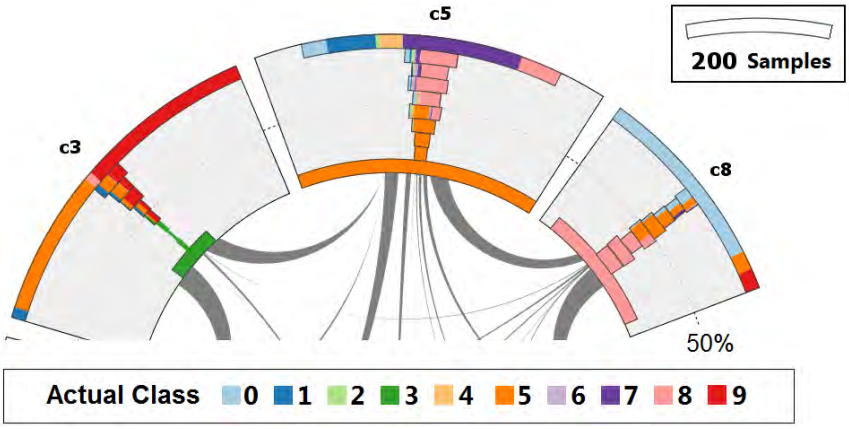


Figure B.4: Filtering and coloring the samples in the confusion wheel, representing the same data as in Fig. B.3b. Only misclassified samples are shown (FPs and FNs), colored by their actual classes.

confusion between a pair of classes and the asymmetry of this confusion, compared with two visually-separate cells in a matrix.

Visualizing additional information about the samples

Instead of coloring the histograms by their classification results, a histogram bar can be alternatively colored by an attribute of the samples aggregated in it. For example, color can be used to compare the classification results against another classifier (Fig. B.8). This shows for which samples the classification improved, worsened, or did not change in the depicted data (Sect. B.5).

B.4.2 Feature Analysis View

Investigating the reason behind certain misclassifications and how to improve them depends heavily on analyzing how the data features are distributed among the affected samples. In particular, it is important to find out if certain features discriminate these samples from correctly classified samples. Therefore, we created a dedicated view that assigns visual primacy to the features by depicting how they are distributed in a selected subset of samples $\hat{S} \subseteq S$. As we did in the wheel view, we split the samples in $\hat{S}$ into the same four groups according to their classification result in a specific class $\bar{c} \in C$ selected by the user. To provide an overview first, we create up to four boxplots below each other for each data feature, showing how its values are distributed in each of the four groups $\hat{TP}_j$, $\hat{FP}_j$, $\hat{TN}_j$, $\hat{FN}_j$. If a group is empty, e.g. no FNs for $\bar{c}$ in $\hat{S}$, no boxplots are created for it. The view shows boxplots of multiple data features ordered in a list (Fig. 1b and C.6g).

Typical feature analysis scenarios involve finding features that separate two main groups among selected samples: $\hat{TP}_j$ from $\hat{FP}_j$, or $\hat{TN}_j$ from $\hat{FN}_j$ (task **T6**). This has two implications on our design: First, we used a minimal version of boxplots, showing the whole value range, the median $Q2$, and the inter-quartile range $[Q1, Q3]$. We do not show outliers as they are

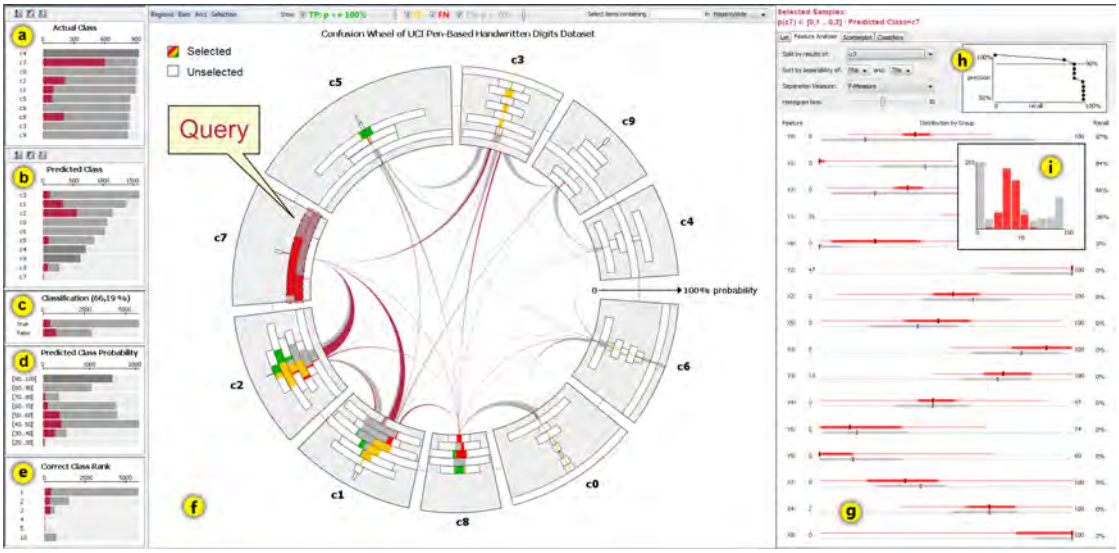


Figure B.5: The interactive exploration environment showing information about classification results, class probabilities, and feature distributions. The *summary charts* show breakdowns of the samples by (a) actual class, (b) predicted class, (c) classification correctness, (d) the probability of the predict class, (e) the probability rank of the actual class. The *wheel view* (f) shows the same data as in Fig. B.4a. Selected samples are highlighted in color. The *feature analysis view* (g) shows summary information data features of the selected samples, broken down according to their classification results, and the selection criteria in natural text. (h) A control panel to rank the features according to their separation power along with a recall-precision curve for possible separation. (i) A histogram of the top ranked feature.

irrelevant for the separation task. Second, and more importantly, we rank the features $f_{1\leq k\leq l}$ by their separation power of two of selected groups X_1, X_2 from the above four groups. Several separation measures can be used for this purpose. One method is to use a significance statistic: for each feature f_k, we compute Welch's t-statistic [34] (which is used for Student's two-sample t-test with unequal sample sizes):

$$t_k = \frac{mean(f_k(X1)) - mean(f_k(X2))}{\sqrt{var^2(f_k(X1))/|X1| + var^2(f_k(X2))/|X2|}} \tag{B.2}$$

We rank the features by the corresponding p-values, computed according to Student's t-distribution. This places features with more significant mean differences between X_1 and X_2 in the top of the list. Such features are more likely to separate the samples in both groups, assuming the values in these groups are normally distributed as in Fig. 1c.

Boxplots show only summary information of feature distribution. To gain more details about it, the user can click on a feature's area, which shows a stacked histogram of its values, depicting breakdown of the samples into the multiple groups described above. It helps in better estimating the separability of the groups by the selected feature.

In many cases, higher mean difference between two groups does not mean better separability. An example is shown in Fig. C.6i, where two groups have relatively closer means, and yet better separability by the respective feature than by other features. This often happens when one of the groups represents combined phenomena like TNs that belong to different actual classes. To account for such cases, we provide alternative separation measures to rank the features. Both χ^2 and K-S statistics [96] are applicable generic measures to compare two empirical distributions without further assumptions. When defining additional classification rules (Sect. B.5.4), it is important to identify *ranges* in feature distributions that have high separation. For this purpose, we use the following measure, based on the histograms h_{1k} and h_{2k} of a feature f_k in X_1 and X_2 respectively:

$$F_k = \sum_{b=1}^{b_h} h_{1k}(b) \cdot \frac{h_{1k}(b)}{h_{1k}(b) + h_{2k}(b)} \tag{B.3}$$

Similar to χ^2, this measure is computed from binned distributions with an adjustable number b_h of histogram bins. Instead of summing up deviations, it sums up the number of X_1 samples in each bin weighted by their retrieval precision, as with the F-measure [115].

To provide an overview of how much separation of X_1 from X_2 is possible using only one of the data features, we create a recall-precision graph in the top of the view (Fig. C.6h). This graph indicates for each precision level to retrieve X_1 , the largest recall rate possible.

In some cases, no single feature provides good separation of the groups. Therefore, we offer a scatterplot view of selected samples $\hat{S}$ in the 2D space of two selected features, to check if these features offer better separation (Fig. 1d). Automated and visual techniques can be employed to recommend scatterplots with best separation [110, 132, 148]. In our implementation, the user selects the scatterplot dimensions from two lists of features, ranked by their univariate separability.

The visual tools described so far show different aspects of classification data. In the next sections we show how these tools are integrated together and describe use cases of our approach.

B.4.3 The Interactive Exploration Environment

Analysis scenarios of classification results typically involve examining different pieces of the information to formulate and test hypothesis about the results, and to introduce improvements. Therefore, we developed an exploration environment that shows these pieces of information at different levels of detail using multiple views that are arranged and coordinated accordingly in the user interface.

Summary views

These views show summary information about the classification results and performance. They assign visual primacy to the classification results, which are shown as secondary information in color in the other views. Each view highlights samples currently under selection. Three bar charts show breakdowns of the samples $s \in S$ by their actual class $l^a(s)$ (Fig. C.6a), predicted class $l^p(s)$ (Fig. C.6b), and classification correctness whether $l^p(s) = l^a(s)$ or not (Fig. C.6c).

In addition, two histograms show breakdowns of the samples by the probability of the predicted class $p_{ij} : l^p(s_i) = c_j$ (Fig. C.6d) and by the rank $r(s_i, l^a(s_i))$ of the actual class $l^a(s_i)$ (Fig. C.6d), where:

$$r(s_i, l_j) = |\{1 \leq j' \leq m : p_{ij'} > p_{ij}\}| + 1 \tag{B.4}$$

Confusion wheel view

This is the central view in the user interface, showing aggregated information in more details than the summary views. Four checkboxes and two sliders at the top of this view enable quickly defining which samples to include. The sliders allow filtering TPs with high probability and FNs with low probabilities, to focus the analysis on more problematic samples to be depicted at a higher resolution (tasks **T3**).

Hovering the mouse over a visual element shows a tooltip with summary information about the samples that it represents (Fig. B.6a). This encompasses, for example, recall and precision in a class, and the number of samples confused between two classes for a chord. More details about selected samples can be obtained in the detail views.

Detail and feature analysis views

The detail views show more information about the samples selected in other views. The top area in these views show a textual description of current selection. A tabular list shows the data features as well as the predicted and actual classes for the selected samples (Fig. B.6b). The feature analysis view shows this information graphically, as explained in Sect. B.4.2. Likewise, the *probability view* shows the class probabilities of the samples as a tabular list or as multiple histograms (Fig. B.6c). The two tabular lists are synchronized: clicking on an sample in one view highlights it and ensures its visibility in both views. This enables a textual examination of the class probabilities of a certain samples. These probabilities are also depicted graphically as a star graph in the wheel view (Fig. B.6a). Also, when possible, a graphical representation of this sample can be shown in a dedicated view (Fig. B.6d).

Interactive Queries on the Samples

By clicking on a bar in the summary views or in the wheel view, the respective subset $E \subseteq S$ of samples is selected (task **T5**). The views are immediately updated to highlight the fractions of bars and chords that represent elements in E. These fractions in the wheel view retain their colors. The rest of the elements become uncolored.

Multiple bars can be selected at once in a histogram in the wheel view. The selection in Fig. C.6 is initially defined by brushing the range $[0.1, 0.3]$ over the radial dimension in c_7. This selects samples s_i with $p_{i,7} \in [0.1, 0.3]$. The selection is refined by excluding predicted samples in c_7, focusing only on TNs and FNs. These samples are highlighted in other sectors to show their classification results in the respective classes. The feature analysis view is updated to show the feature distributions among these samples.

The samples confused between two classes can be selected by clicking on the respective chord (task **T4**). This allows examining how these samples are distributed in the probability histograms of both sectors. The selection in Fig. B.6a is defined by clicking on the chord between c_2 and c_7. The views are updated to show the class probabilities and feature values of the samples in E. These samples can be examined individually by clicking on an item in these views (Fig. B.6c). Finally, samples can be further selected based on their features or other attributes by selecting a specific value range in the respective view [1].

The subsets that correspond to the above-mentioned selection possibilities can be combined interactively using set operations as in [5]. Specific keyboard modifiers allow specifying if the newly brushed elements should be added to, intersected with, or subtracted from the existing selection. This allows defining highly-expressive visual queries to select samples based on their classification results and probabilities in each class, and on their actual and predicted classes. For example, by clicking on c_3 in "predicted class" summary chart all positive samples in this class are selected (both TPs and FPs). This selection can be refined to TPs only by clicking on the respective bar in "actual class" while in set intersection mode. Also, certain FPs such as confusions with c_5 and c_9 can be filtered out by clicking on their bars in this chart while in set exclusion mode. We show in the next section how interactive selection of the samples supports several analysis scenarios of probabilistic classification data.

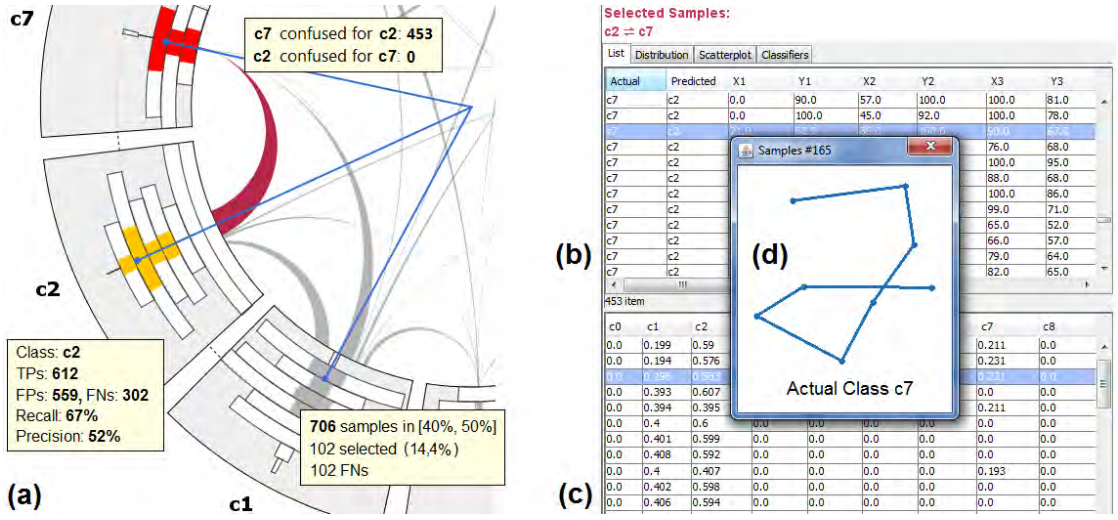


Figure B.6: Obtaining details about elements interactively: (a) selecting samples confused between c_2 and c_7. The list view show their (b) features and (c) class probabilities. One sample is selected for inspection by showing its graphical representation (d). Its class probabilities are highlighted both in the list view, and in the wheel view using arrows.

[1] The supplementary video illustrates some of the interaction possibilities

B.5 Usage Scenarios

We demonstrate how our tools can be applied to analyze and improve classification results of the UCI "pen-based" dataset introduced in Sect. B.4. For this purpose, we train several classifiers using the raw features [2] of 100 randomly-selected samples using the RapidMiner data-mining software [136] (formerly YALE [101]). We first show how interaction allows quick inspection of misclassified data. Then we show how the confusion wheel provides insight into classifier behavior, and enables comparing misclassified samples between two classifiers. Finally, we show how our tools help in defining additional classification rules to correct misclassified samples in a generalizable way. Besides demonstrating the standard features of our system, problem-specific features are introduced to support the last two use cases. Further examples with different data sets and classifiers are available at `http://www.cvast.tuwien.ac.at/ConfusionAnalysis/`

B.5.1 Inspecting Misclassified Samples

The rich possibilities to select subsets of samples using the interactive exploration environment allow quick inspection of certain samples, e.g., to analyze the reason behind certain confusions. In Fig. B.6, elements confused between c_2 and c_7 are selected by clicking on the respective chord. Inspecting these samples illustrates that people write the same digits in different ways, which requires increasing the training sample to match against using *k*-NN classifiers.

In many cases, erroneous labels or noisy data features are the reason behind classification errors. Using our detail views we were able to identify such cases in the UCI dataset. One example is a c_5 digit confused for c_8 because it is written in east Arabic numerals which have different shapes than Arabic numerals. Another example was a noisy sample which does not resemble any digit. Identifying and isolating such samples is important to introduce effective design improvements and to accurately evaluate and compare different classifiers.

B.5.2 Analyzing Classifier Behavior

Visualizing the probability histogram for each class and coloring these histograms by classification results reveal several patterns that explain the behavior of the classifier. Fig. B.7 shows this information for class c_5 using three different classifiers. With Neural Networks (NN), the classification accuracy increases proportionally with the probability of the predicted class (Fig. B.7a). This does not always apply to a Naïve Bayesian (NB) classifier, where some classes showing the opposite trend (Fig. B.7b). Also, though it varies between classes, the overall classification sharpness was higher for NB than NN, with 88.9% of all samples classified with $\geq 90\%$ probability (Fig. B.3b), as opposed to 50.7% with NN. *k*-NN classifiers exhibit up to k peaks in the histograms at equidistant locations, when an appropriate number of bins b is used (Sect. B.4.1). This is because *k*-NN classifiers perform weighted majority voting among the labels of the k nearest training samples to the samples being classified. If all k classifiers agree on the label c_j for s_i, p_{ij} is equal (or very close to) 1 and the sample belongs to the outermost peak in the

[2] For reproducibility and illustration purposes, we did not consider computing any additional features that could improve the performance.

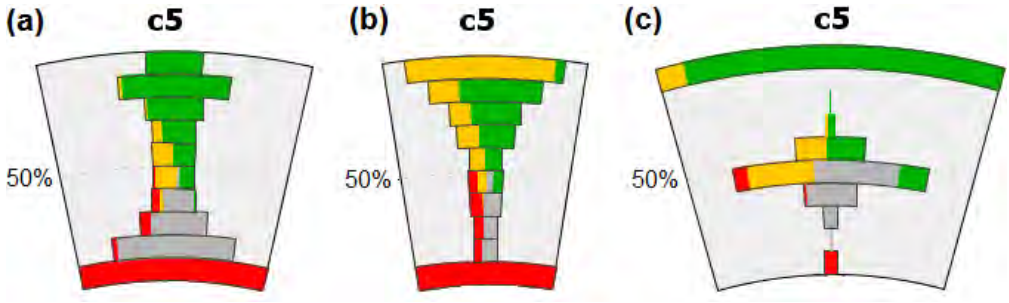


Figure B.7: Classification results in c_5 using (a) neural networks, (b) a naïve Bayesian classifier, and (c) a *k*-NN classifier with $k = 2$. TNs in the bottommost bars are filtered out. Individual histogram scales are used.

histogram. If none of the classifiers agree, the label c_j of the closest training sample is predicted but with low probability, and the sample hence belongs to the innermost peak in the histogram of c_j. In Fig. B.7c there are two peaks, as k equals 2. In Fig. 1a and Fig. C.6f, there are up to five peaks per sector, as k equals 5.

B.5.3 Comparison with Another Classifier

Classifier designers typically analyze the effect of using a different classification algorithm or changing certain parameters on the results. Besides a holistic measure of classification rate improvement, they often want to gain insight about the samples whose classification improved by this change, and the ones that worsened. A typical example is analyzing how the classes vary in their improvements, using two-sided bar charts that depict improved and worsened samples for each class in different directions. Another example is analyzing which class confusions increased or decreased by the changes, using a suited matrix representation. These representations discard available information on class probability which offers new ways to improve the performance. To address this limitation, we offer a mode to color the samples in confusion wheel by their improvement status as illustrated in Fig. B.8. The data is classified using a *k*-NN algorithm with $k = 1$ and $k = 3$. The histograms show the class probabilities computed with $k = 3$. TNs are not shown, as they are irrelevant for comparison on class level. Dark blue indicates misclassified samples in the depicted data that would improve when $k = 1$. Dark red indicates correctly classified samples that would worsen when $k = 1$. It is noticeable that $k = 1$ performs better than $k = 3$ as there is more dark blue than dark red (overall 6% improvement). We investigated the reason for that by checking samples that improved or worsened. Fig. B.8a illustrates an example of a sample whose nearest neighbor is the correct class, but the 2nd and 3rd nearest are not. In this example $k = 1$ succeeds while $k = 3$ fails. Fig. B.8b shows the opposite case: 2 out of 3 nearest neighbors have the correct labels, making $k = 3$ succeeds and $k = 1$ fails. In both cases, the sample is close to the outer boundary in c_2 as two of the three nearest neighbors agree on its label.

Fig. B.8c shows an interesting case which resembles Fig. B.8a, with the only difference that the 2nd and 3rd neighbors are significantly far from the sample. This makes the classifier less

confident about their votes, and hence predicting the answer at lower probability. Except for one sample, all 480 samples that fall in this probability range in this class would either improve or stay the same when $k = 1$. This suggests adding a rule to re-classify these samples, as we show next.

B.5.4 Defining Post-Classification Rules

In some cases, an existing classifier cannot be refined internally due to the lack of source code or expertise. Post-classification rules are one way to improve the performance in such situations by handling specific cases [157], incorporating domain knowledge [33], and rectifying systematic errors. Such rules are usually defined over the data features and are usually easy to understand and adapt especially if they are defined by domain experts. In case of probabilistic classification, a rule can specify Boolean conditions $q(s_i)$ on the class probabilities p_{ij} or ranks (Eq. B.4) of the samples s_i, in addition to conditions on their features $f_k(s_i)$. Both types of information are available at runtime as they do not involve the actual labels $l^a(s)$.

We consider three rules for correcting classification errors:

- **Correcting false negatives:** This rule intends to correct FNs of class c_j by replacing their predicted classes with c_j:

$$R_{FN_j} : q_1(s_i) \wedge \ldots \wedge q_k(s_i) \wedge l^p(s_i) \neq c_j \Rightarrow l^p(s_i) \leftarrow c_j \tag{B.5}$$

  Samples that satisfy conditions $q_1..q_k$ are post-classified as c_j.
- **Correcting false positives:** This rule intends to correct FPs of class c_j by replacing their predicted classes with the 2nd guesses:

$$R_{FP_j} : q_1(s_i) \wedge \ldots \wedge q_k(s_i) \wedge l^p(s_i) = c_j \Rightarrow l^p(s_i) \leftarrow c_{j'} : r(s_i, c_{j'}) = 2 \tag{B.6}$$

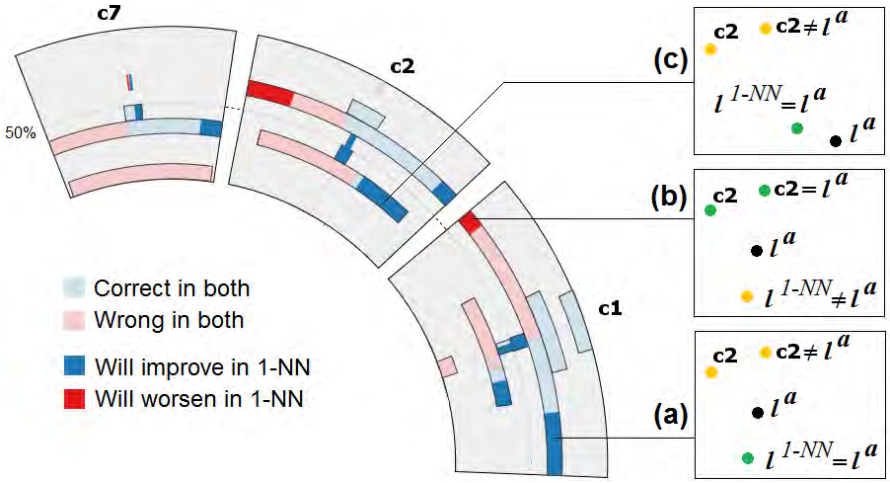


Figure B.8: Comparing the results of k-NN classifiers with $k = 1$ (encoded in color) against $k = 3$ (defining the histograms).

It post-classifies a potential FP that satisfies its premise as the class $c_{j'}$ ranked 2nd for this sample (Eq. B.4).

- **Using another classifier:** This rule intends to re-classify certain samples by using another classifier Cl_z:

$$R_{Cl_z} : q_1(s_i) \wedge \ldots \wedge q_k(s_i) \Rightarrow l^p(s_i) \leftarrow Cl_z(s_i) \tag{B.7}$$

It post-classifies a sample s_i that satisfies its premise as the class $Cl_z(s_i)$ predicted by Cl_z.

Our visual tools support in defining rules of the above types and in testing their actual improvement. For this purpose, the data set should be split into two parts: (1) training data that are loaded in the visualizations and used in defining the rules, and (2) validation data that are used to assess the actual improvement on unseen data. This is important to avoid dataset bias which occurs when defining rules that overfit the training data and fail to generalize to unseen data. Except for Fig. B.3, the visualizations depicted in this paper use 80% of the UCI data introduced in Sect. B.4. In the following we illustrate how potential improvements can be visually identified, and how the respective rules can be defined and validated on the remaining 20% of the data.

To improve misclassified samples in a class c_j, the respective rule should define conditions on the samples that include as much of these samples as possible and in the same time exclude correctly-classified samples. This is important as applying the rule on the latter samples would worsen their classification. The wheel view gives an overview on how misclassified samples c_j are distributed according to their probabilities. This makes it easy to spot probability ranges in c_j having a significant number of these samples that are likely to improve by one of the rules. Typically, these samples interfere with correctly classified samples. In particular, the interference of TPs (green) with FPs (yellow) requires separation using further conditions on the data features, before applying R_{FP_j}. This interference is usually larger in the outermost bin(s) that are dominated by TPs, which suggests excluding these bins when defining this rule. Similarly, the interference of FNs (red) with TNs (grey) requires separation in order to apply R_{FN_j}. Also the innermost bin(s) need to be excluded when defining this rule, as their probability range is highly dominated by TNs.

To separate the interference in a potentially improvable probability range Q_{c_j} in c_j, the user selects the samples in this range using brushing. The feature analysis view lists possible features that offer good separation, as explained in Sect. B.4.2. After inspecting the feature histograms, the user can select a feature f_k to define an improvement rule by double clicking on its histogram. This opens a dialog box which shows the histogram in higher resolution and allows selecting a specific value range Q_{f_k} for f_k (Fig. B.9). The user selects a range that contains the majority of samples that need improvements (FPs or FNs) and excludes as much of the other samples as possible. The dialog also allows specifying which rule to apply to samples that fall both in Q_{c_j} and Q_{f_k}. The selected rule is externalized in text and applied to such samples both in the training dataset loaded in the visualization, and in the test dataset. The results in both cases are reported as the absolute number of samples that improved and worsened, and the total improvement on the classification rate. Changing the feature range Q_{f_k} automatically updates the results, to assist the user in choosing a robust range that performs well both in training and test datasets.

As example, in Fig. C.1a, the analyst notices a large number of FNs in c_7. She selects the respective probability range in c_7 (Fig. C.6f) and finds that feature y_8 offers good separation of these FNs from the TNs in the value range $[20\%, 60\%]$ (Fig. C.6i). Therefore, she creates the following rule:

$$(0.1 \leq p_{i7} \leq 0.3) \wedge (20 \leq x_8(i) \leq 60) \Rightarrow l^p(s_i) \leftarrow c_7 \tag{B.8}$$

This rule improves 591 and worsens 13 samples in the loaded data, yielding a significant total improvement rate of 6.57%. Similar results apply to the testing data (141 improved, 3 worsened, 6.28% total rate) making the analyst accept this rule. She continues further to investigate the large number of FPs in c_1 by selecting the probability range $[30\%, 80\%]$ and restricting the selection to samples with $l^p(s) = c_1$. She checks the features for separation but notice interference between FPs and TPs, even with the features with most separation power (Fig. B.9a). She selects the small range with as few TPs as possible, and applies rule R_{FP_1} which achieves 1.11% overall with 136 improved and 36 worsened samples in the training dataset. She rejects this rule and searches for more robust rules to improve these samples.

The scatterplot view allows identifying if two features can in combination achieve a good separation of interfering sample groups. As example, Fig. 1d illustrates how two features separate about 76% of FNs in c_8 that lie in probability range $[10\%, 30\%]$ from the TNs, with only 13 TNs unseparated. Reclassifying these samples with R_{FN_j} yields 3.5% and 3.4% improvements on the training and test datasets. Dedicated algorithms are needed to rank the pairs of features by their separation power, and to recommend optimal region separations in a specific scatter plot. For the purpose of this use case, a manual search for such features is sufficient to illustrate the importance of visual inspection to assess their separation power.

Besides searching for separating features, it is possible to improve certain misclassifications using the results of another classifiers. As example, it is evident in Fig. C.1a that the results for c_2 involve a large number of mixed misclassifications (FNs and FPs) especially in the probability range $[30\%, 70\%]$. We provide a separate view to check how other classifiers perform on these samples by selecting them (Fig. B.9b). This reveals that neural-networks-based classifier (NN) yields 7.31% improvement rate if applied to these samples, which suggests creating the following post-classifying rule (Eq. B.7):

$$0.3 \leq p_{i2} \leq 0.7 \Rightarrow l^p(s_i) \leftarrow l^p_{NN}(s_i) \tag{B.9}$$

Such combinations of results from multiple classifiers has been extensively researched in pattern-recognition and machine-learning literature [89, 149, 172]. Many of these techniques apply combination heuristic such as weighted majority voting in a holistic way to all samples. We illustrated that declaratively restricting such heuristics to certain samples yields better improvements, as this avoid impacting correct classifications among the remaining samples. Using visual inspection, we were able to outperform automated techniques for combining multiple classifiers such as majority voting (Fig. B.9b).

Post-classification rules should be defined carefully to avoid over-fitting the data. First, the testing dataset should be representative and of sufficient size to warrant generalization. Second, the conditions used in these rules should be based on probability and feature ranges that have a sufficient number of samples to avoid creating rules specific to the training dataset. In fact,

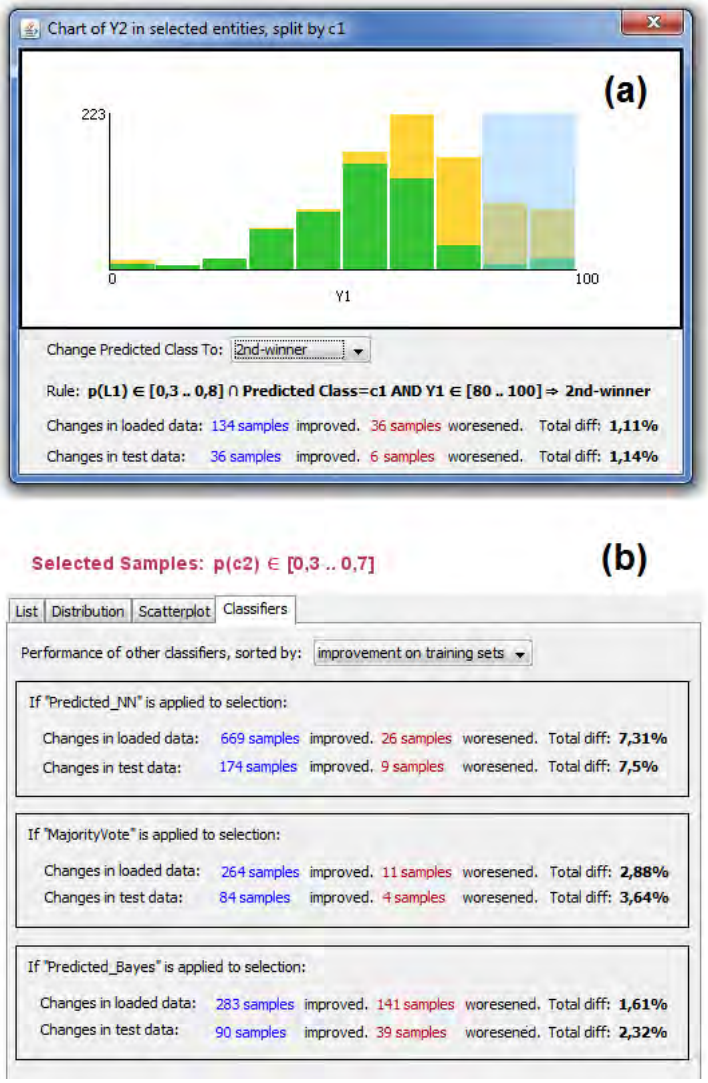


Figure B.9: Defining post-classification rules: (a) checking separability between FPs (yellow) and TPs (green) according to a feature, (b) checking the performance of other classifiers on selected samples.

the distributions depicted in the probability histograms tend to be invariant among random subsets of sufficient sizes, if extreme values are discarded and avoided. Finally, defining several rules increases the overlap between their premises, the conflict in their actions, and the risk of over-fitting in general. It is important to select a small number of rules that exhibit robust improvement results, high precision, and few overlaps with each other. In general, classification errors should ideally be solved by improving the classification algorithm when possible, with help of the insights gained by the interactive visualizations.

B.6 Discussion

In this section we discuss the advantages and limitations of our tools and compare them with other techniques for visualizing probabilistic classifiers. We also report feedback and observations from classification experts and practitioners.

Scalability: The use of aggregated representations makes our tools highly scalable with the number of samples. For example, the histograms in the confusion wheel were able to handle datasets containing tens of thousands of samples. Furthermore, the filtering possibilities presented in Sect. C.2.3 enable focusing on fine details contained in small subsets of these samples. Likewise, the boxplots and histograms in the feature analysis view scale well with the number of samples. Stacking the boxplots allow depicting summary information about 15-20 features at once. This is sufficient for our purposes, thanks to feature ranking which interactively places the most relevant features for the current analysis context at the top.

To ensure enough visibility of the information in each class, up to 20 classes can depicted at once as sectors in the wheel view. This limit is feasible for a wide range of classification problems that do not require a larger number of classes. In case of larger number of classes, a subset of them can be selected manually or automatically to be shown at once, such as the subset with the highest confusions between its classes.

Handling imbalanced data: Sometimes, classification data exhibit skewed distributions of the samples to the classes. This causes classes having large number of positives to occupy the majority of display area, possibly obscuring fine details in smaller classes. To handle such cases is to possible to make the sectors of equal sizes, and stretch the histogram to fit in these sectors using individual scaling factors. Arcs representing the same amount of samples can be drawn outside the sectors using different scales to indicate these scaling factors. Although this hinders comparing the histograms in absolute values, the shapes of the distributions and the proportions of misclassified samples are still comparable across the classes. Such relative comparisons are usually more relevant in analyzing the results than comparing absolute values between classes of significantly different sizes. Similarly, the class confusions can be normalized by the total number of samples in the respective classes. These relative confusions can be indicated in color or by adjusting the chord thicknesses to show relative instead of absolute confusions.

Some classifiers compute relatively low probabilities for the winning class, such as ones based on fuzzy logic. This limits non-empty histogram bars to the inner bins only, which lowers the visual resolution. It is possible in such cases to redefine the bins to cover the effective probability range at higher resolution and make the histogram span the whole sector area.

Comparison with previous work: Our tools combine different pieces of information that have been addressed individually in previous work, as presented in Sect. B.2. Compared with existing techniques that visualize class probabilities such as Class Radial Visualization (Fig. A.2), the added-value in using the wheel metaphor lies in (1) aggregating the samples to analyze their probability distributions and to avoid clutter and ambiguity issues caused by individual points [134], (2) using color to show classification results of the samples next to each other and

to reveal their separability, and (3) visualizing class confusions in a compact layout, thanks to the radial arrangement. This provides a rich overview of classification results that was not possible in previous techniques and enables new insight and tasks as we illustrate in the usage scenarios. Confusion matrices are better suited than the chords to gain more precise information about class confusions. Nevertheless, the chords reveal groups of classes having higher mutual confusions, emphasize the asymmetries in these confusions, and enable relating them to respective class probabilities.

The area-based nature of the histograms offers several possibilities for selecting and highlighting certain subsets of samples in confusion wheel. This is essential in a coordinated-multiple-view environment in order to formulate queries on the samples and to gain detailed insight into them in the feature analysis view. This environment enables a novel integration between probability-based and feature-based representations of classification data.

The visual analysis methods we propose are based on familiar visual representations such as boxplots and stacked histograms, as well as on the wheel metaphor [4]. This metaphor is originally designed to visualize associations between entities and categories, and similarities between categories. By treating samples as entities and classes as categories, we were able to adapt this metaphor for classification data: Instead of associations and similarities, we compute probabilities and two-way confusions. Also, we introduced several changes to the visual encoding and different interactions to address the characteristics of classification problems.

Expert feedback: We analyzed classification data comprising about $40,000$ labeled samples provided by our industry partners. The data were classified in 10 classes using a fuzzy rule-based system designed by domain experts. We refined our design iteratively based on feedback from these experts, and on what information they wanted to analyze. After we explained the visualizations we created for their data, they were able to identify issues with their classification rules. For example, one class had a large number of FNs that were highly concentrated in the middle of the respective sector. Analyzing the reason for that revealed that the respective classification rule was assigned relatively low weight by the domain experts. Also, examining the features for selected FPs in a class revealed a rule that uses an inappropriate feature that was mistaken for the correct one: The values of this feature among these FPs should not appear among actual samples of the class, according to domain experts. These experts were able to refine their system accordingly. We did not have the possibility to quantify the improvement.

The informal feedback from our industry partners and from five other machine-learning experts confirms that our tools offers both a good overview of classification results and detailed information on demand. However, our visual metaphors need to be learned with enough examples and explanations before they can be interpreted correctly: One domain expert, asked for clarification on what the histograms in the wheel view mean, about 30 minutes after we started presenting our findings. To avoid such misunderstanding, the metaphor should be introduced part by part with sufficient examples, before discussing insights. In fact, our system is feature laden, and needs extensive learning of how these features work together. One machine-learning expert requested showing separate arcs to encode class confusions in both directions. Also, two machine-learning experts did not encourage using post-classification rules in general, as they could encourage over-fitting the data. They suggested using the insight gained in improving the

classifier design instead.

B.7 Conclusion

The availability of class probabilities enables new possibilities to analyze the performance of probabilistic classifiers, beyond comparisons between predicted and actual classes. Common visual representations such as confusion matrices and ROC curves ignore class probabilities by assigning visual primacy to classification error in terms of false positives, false negatives, or class confusion. Assigning visual primacy to class probabilities or to the data features enables analyzing their influence on classification performance and performing further analysis tasks related to the data. We proposed a representation of probabilistic classification data by showing the probability distributions as stacked histograms in a radial layout and by coloring these histograms by the classification results of the samples. We showed how this representation lends itself to rich interactions to select samples based on their probabilities, and to perform further analysis of these samples based on their data features. We proposed intertwined automated and visual methods to analyze these features in a dedicated view and to rank them according to their separation power between true and false classifications among the selected samples. We presented several analysis scenarios that are possible using our visual tools. These include visual inspection and comparison of classification results, identifying performance problems, and interactive definition of post-classification rules to improve misclassified samples *a posteriori*. We demonstrated by that how exploratory analysis can reveal relevant patterns and correlations in classification data that are difficult to specify and identify automatically, and are usually compromised in holistic analysis methods. These insights are essential to introduce effective improvement to the classifier design that reduce the classification error in a generalizable way. Our future work aims to provide visual means to compare and combine the results of several classifiers, to support analyzing a large number of classes, and to explore hierarchical and multi-label classification results by means of similar interactive visualization methods.

Acknowledgement: We thank Peter Filzmoser from TU Vienna for feedback on separation measures, Colin Johnson from the University of Kent for feedback on confusion wheel, and Bilal Esmael and Arghad Aranout from the TDE Data Engineering for cooperation and discussions. This work was supported by the Austrian Federal Ministry of Economy, Family and Youth via CVAST, a Laura Bassi Centre of Excellence (project no. 822746).

CHAPTER C

Reinventing the Contingency Wheel: Scalable visual analytics of large categorical data

Appears in IEEE Transactions on Visualization and Computer Graphics, 18(12):2849-2858, 2012.

Abstract: Contingency tables summarize the relations between categorical variables and arise in both scientific and business domains. Asymmetrically large two-way contingency tables pose a problem for common visualization methods. The Contingency Wheel has been recently proposed as an interactive visual method to explore and analyze such tables. However, the scalability and readability of this method are limited when dealing with large and dense tables. In this paper we present Contingency Wheel++, new visual analytics methods that overcome these major shortcomings: (1) regarding automated methods, a measure of association based on Pearson's residuals alleviates the bias of the raw residuals originally used, (2) regarding visualization methods, a frequency-based abstraction of the visual elements eliminates overlapping and makes analyzing both positive and negative associations possible, and (3) regarding the interactive exploration environment, a multi-level overview+detail interface enables exploring individual data items that are aggregated in the visualization or in the table using coordinated views. We illustrate the applicability of these new methods with a use case and show how they enable discovering and analyzing nontrivial patterns and associations in large categorical data.

keywords: Large categorical data, contingency table analysis, information interfaces and representation, visual analytics.

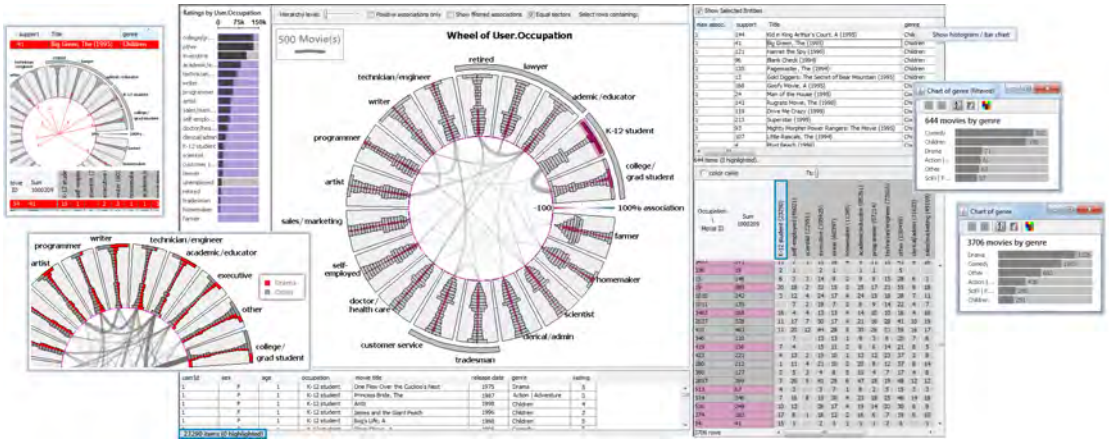


Figure C.1: Contingency Wheel++ uses complementing visual representations and a multi-level overview+detail user interface to enable rich exploratory analysis of large categorical data. The example above shows information about 1 million user ratings on 3706 movies.

C.1 Introduction

Many problems in scientific domains such as medicine, biology and pharmacology, as well as in business domains such as retail and logistics require analyzing associations between categorical variables. For example, a movie retailer might be interested in associations between movies and users based on sales data with the goal of optimizing marketing strategies. The discrete nature of categorical data and their lack of an inherent similarity measure pose significant challenges to the fields of information visualization [13] and data mining [171]. Contingency tables (also known as crosstabs) are a common way to summarize categorical data as a first step of analysis. A two-way contingency table is an $n \times m$ matrix that records the frequency of observations f_{ij} for each combination of categories of two categorical variables. Many data analysis frameworks such as KNIME [16], WEKA [54] and R [116] offer possibilities to create and analyze contingency tables. One of the best-known statistical tests for the overall association (or independence) between two categorical variables is Pearson's χ^2 test [119]. It assesses the significance of associations between the categories of the two variables. However, it does not provide information about how single pairs of categories are associated.

Several visualization methods were developed to analyze associated categories in contingency tables. As we discuss in Sect. C.4, these methods are designed to handle rather small tables having few categories. However, often much larger contingency tables need to be analyzed, which poses a problem to these methods. Figure C.2a shows large categorical data from the MovieLens data set [53]. It contains about one million user ratings on movies. For each user, it provides his or her occupation, sex, and age group, and for each movie, its release date and genres. Examples for tables extracted from this data set are:

- A 3706×21 table which counts for each movie, how many times it was rated from users

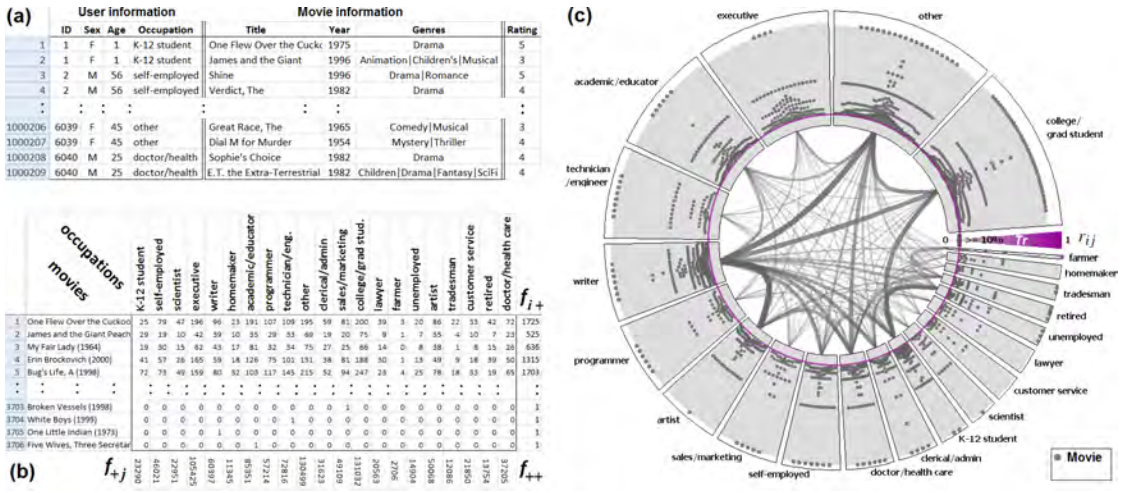


Figure C.2: (a) Categorical variables of the MovieLens data set [53] showing about one million user ratings on 3706 movies, (b) the contingency table of the variables "movie title" and "user occupation", (c) the Contingency Wheel of the table in (b): Sectors represent occupations and dots represent movies positively associated with them. Thicker arcs show which occupations share more movies highly associated with both of them.

of each occupation (figure C.2b).

- A 6040×17 table which counts for each user, how many times he/she rated movies from each genre (figure C.6d).

The Contingency Wheel [6] has been introduced as an interactive visual method for exploring positive associations in asymmetrically large tables. The column categories are visualized as sectors of a ring chart and the table cells are depicted as dots in these sectors (figure C.2c). The dot for cell (i, j) is placed in sector i at a radial distance from the ring's inner circle proportional to the strength of association r_{ij} between row i and column j. A layout algorithm calculates the angular positions of the dots in each sector to reduce occlusion. It copes with a large number of rows by visualizing only the cells that represent significant associations r_{ij}, determined by adjustable thresholds. An arc is drawn between two sectors if one or more rows have dots in both sectors. This arc is thicker if more such rows exist and if their dots represent higher associations with both sectors. User interaction enables analyzing different types of associations in large tables.

Scalability is one of the major challenges visual analytics aims to address [152]. The wheel metaphor explained above has several shortcomings which degrade its readability and scalability, especially with large and dense tables (Sect. 3.7.3). In this paper we propose Contingency Wheel++: new visual analytics methods that tackle the issues of the original wheel. Our methods (described in Sect. C.2) address its computational component, visual representation and interactive interface, and intertwine these three components to enable scalable analysis of categorical data. The new methods encompass:

- Automated methods: a new association measure results in a better distribution of the dots to sectors of different sizes. This is important when analyzing large tables that often exhibit high skewness in the distribution of their frequencies.
- Visualization methods: a frequency-based abstraction of the dots eliminates overlapping which allows showing all the cells, instead of just small subsets thereof. This enables analyzing and querying both positive and negative associations.
- Interactive exploration environment: an overview+detail interface allows exploring individual items aggregated in the visualization or in the table, and analyzing their attributes.

In Sect. C.3 we present a use case to illustrate how our new methods can be used to explore the MovieLens data set. We show how nontrivial patterns and associations in the data can be discovered. In Sect. C.4 we compare our approach with other methods for visualizing categorical data and elaborate on its scalability.

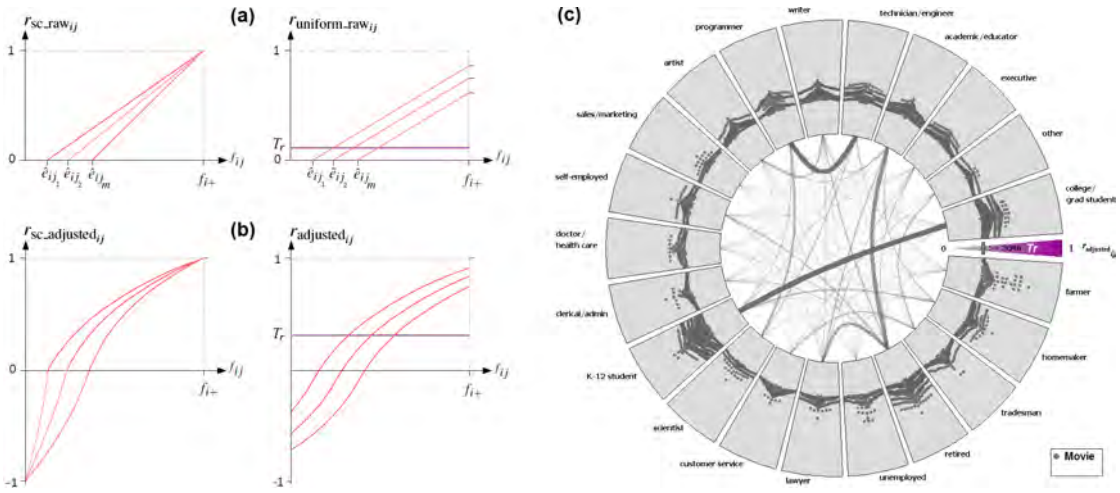


Figure C.3: (a) raw residuals and (b) adjusted residuals plotted as a function of f_{ij} for different values of f_{+j} with both nonuniform- (left) and uniform scaling (right), (c) the same data plotted in figure C.2c using uniformly-scaled adjusted residuals instead of raw residuals (with $T_r = 30\%$ and $T_s = 1$).

Contingency Wheel++ [1] improves both on the readability and on the scalability issues mentioned above by employing visual analytics methods as presented in the next section.

C.2 Contingency Wheel++

In the following, f_{ij} denotes the frequency in cell (i, j), $f_{i+} = \sum_{j=1}^{m} f_{ij}$ and $f_{+j} = \sum_{i=1}^{n} f_{ij}$ are the marginal row- and column frequencies, and f_{++} is the sum of all table frequencies

[1] A prototype implementation of Contingency Wheel++ is available at http://www.cvast.tuwien.ac.at/wheel

(figure C.2b). We first address the data mapping employed by Contingency Wheel++ (Sect. C.2.1). Then we propose a frequency-based visual representation which abstracts the dots (Sect. C.2.2). In Sect. C.2.3 we show how an interactive visual interface integrates additional table views to bridge the gap between the data representation and the visual representations and to support a flexible visual exploration process.

C.2.1 Mapping Frequencies to Associations

The main goal of Contingency Wheel++ is to reveal how the row categories of a contingency table are associated with its column categories. For this purpose, it uses a statistical measure r_{ij} that computes the association between row i and column j based on f_{ij} and takes value in the range $[-1, 1]$. This measure is usually based on statistical residuals between the actual frequency f_{ij} and expected frequencies $\hat{e}_{ij}$. The frequency in cell (i, j) predicted under the null hypothesis H_0, i.e., assuming no association, is [143]:

$$\hat{e}_{ij} = \frac{f_{i+} \cdot f_{+j}}{f_{++}} \tag{C.1}$$

If $f_{ij} = \hat{e}_{ij}$ holds for cell (i, j), its share f_{ij}/f_{i+} of the marginal row frequency is equal to the column's share f_{+j}/f_{++} of all table frequencies. This means that row i is neither positively nor negatively associated with column j, and corresponds to a zero association value $r_{ij} = 0$. Cells with $f_{ij} > \hat{e}_{ij}$ indicate a positive association between row i and column j. Statistical residuals r_{ij} can be used to quantify this association. They can be designed to incorporate a priori information about the data and their distribution. In the following we describe the originally-used residuals and our improvements on them.

Raw residuals

The association measure used originally by the Contingency Wheel is based on raw residuals $(f_{ij} - \hat{e}_{ij})$ [6]. To generate association values $r_{ij} \leq 1$, the raw residual for cell (i, j) is divided by the maximum value it can take $(f_{i+} - \hat{e}_{ij})$:

$$r_{\mathrm{sc\_raw}_{ij}} = \frac{f_{ij} - \hat{e}_{ij}}{f_{i+} - \hat{e}_{ij}} \tag{C.2}$$

This measure maps frequencies linearly to association values (figure C.3a-left). The maximum association $r_{ij} = 1$ is reached when all cells of row i have zero frequencies except for cell (i, j). For such a row, only one dot is created on the outer boundary of sector j. A cell with $r_{ij} = 0$ creates a dot on the inner boundary of sector j (assuming no thresholds). Cells with negative associations are ignored. The above-mentioned normalization is not uniform with respect to the columns: For row i, different scaling factors are used in different columns, because the expected frequency $\hat{e}_{ij}$ is larger for columns with larger f_{+j}. This makes better use of the sector area for revealing the distribution of dots along the radial dimension. Also, rows i that are fully associated with column j ($f_{ij} = f_{i+}$) can be easily found as dots at the outer boundary. However, the different scaling factors result in a bias especially when f_{+j} varies largely between

sectors. This impacts the comparison of associations between different sectors and reduces the expressivity of the arcs. A uniform scaling factor for all columns can be used instead:

$$r_{\text{uniform\_raw}_{ij}} = \frac{f_{ij} - \hat{e}_{ij}}{f_{i+}} \tag{C.3}$$

Figure 3a-right shows how this scaling maps frequencies to associations. For cells with $f_{ij} = f_{i+}$, Eq. C.3 evaluates to $1 - f_{+j}/f_{++}$ which is independent of i. Such cells are hence mapped to the same radial distance within a sector (figure C.2c). The sectors are scaled by their marginal frequencies. Sectors with larger f_{+j} values attract more dots than sectors with smaller f_{+j} values, due to an inherent statistical bias that raw residuals suffer from (even with uniform scaling).

Adjusted residuals

Standardized Pearson residuals [143] avoid the bias of raw residuals by adjusting the variance of the r_{ij} values to $N(0, 1)$:

$$r_{\text{pearson}_{ij}} = \frac{f_{ij} - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij} \cdot (1 - f_{i+}/f_{++}) \cdot (1 - f_{+j}/f_{++})}} \tag{C.4}$$

We use a logarithmic scale for the visual mapping of these residuals to better reveal their distribution along the radial dimension (where cte is a constant computed from the table to ensure $-1 \leq r_{ij} \leq 1$):

$$r_{\text{adjusted}_{ij}} = \frac{\text{sgn}(r_{\text{pearson}_{ij}})}{cte} \cdot \ln\left(1 + \left|r_{\text{pearson}_{ij}}\right|\right) \tag{C.5}$$

Figure C.3b-right, shows how this measure maps frequencies to associations. Figure C.3c shows the same data as in figure C.2c using $r_{ij} = r_{\text{adjusted}_{ij}}$ with $T_r = 30\%$ and with equal sectors. The dots are distributed more uniformly among the sectors. This results in arcs that suggest other similarities between occupations. The logarithmic scale amplifies smaller raw residuals, giving them more visual prominence. This potentially generates more dots, and hence a higher value for T_r is needed to filter out insignificant associations. Cells with $f_{ij} = f_{i+}$ are mapped to different radial distances in sector j, depending on f_{i+}. This makes the arcs more robust to changes in T_s since rows with smaller f_{i+} values contribute less to the arcs. On the other hand, these cells are somewhat difficult to locate. The following nonuniform scaling stretches r_{ij} to the range $[-1, 1]$:

$$r_{\text{sc\_adjusted}_{ij}} = \frac{r_{\text{adjusted}_{ij}}}{\max\left(s_{ij} \cdot r_{\text{adjusted}_{ij}|_{f_{ij}=f_{i+}}}, s_{ij} \cdot r_{\text{adjusted}_{ij}|_{f_{ij}=0}}\right)} \tag{C.6}$$

where $s_{ij} = \text{sgn}(r_{\text{adjusted}_{ij}})$ and $r_{\text{adjusted}_{ij}|_{f_{ij}=x}}$ is the value $r_{\text{adjusted}_{ij}}$ would take if $f_{ij} = x$. Figure C.3b-left depicts how this scaling maps frequencies. As can be seen, rows with $f_{ij} = f_{i+}$ are always mapped to the largest radial distance. Also, if the visualization can include negative associations, rows with $f_{ij} = 0$ are always mapped to the lowest radial distance. Nonuniform scaling, however, re-introduces a small bias in the associations, toward columns with larger f_{+j}.

C.2.2 Visualizing the Contingency Table

The visualization aims to reveal how the row categories of a contingency table are associated with its column categories, based on the association measure used. Our new visual representation makes use of the advantages of uniformly adjusted residuals (Sect. C.2.1). It provides a clearer and more intuitive visualization of the table, as compared to the original wheel design [6]. Moreover, depending on the user's choice, it enables showing all associations or positive associations only as we describe in the following subsections.

Visualizing columns

Like in the original wheel metaphor, columns are drawn as sectors of a ring chart. The main difference is that they are drawn with equal size. This has several advantages: First, this is in accordance with the fact that adjusted residuals result in a more uniform distribution of the cells to the sectors. Second, by using a frequency-based representation (Sect. C.2.2), the distribution of the associations can be compared between different sectors. Third, the arcs become evenly distributed in the central area, unlike the arcs in figure C.2c which overlap more near small sectors. Finally, column categories are treated equally from a visual point of view, in the same way as the dimensions of a star plot [56]. This simplifies the visualization and eliminates confusion about the meaning of different sector sizes. The information of different column marginal frequencies f_{+j} is conveyed in a linked bar chart (Sect. C.2.3). Incorporating it in the wheel representation would not contribute to the goal of Contingency Wheel++, i.e., to explore associations.

Visualizing row-column associations

The radial dimension of the ring chart linearly encodes the association values r_{ij} computed by one of the association measures. The outer boundary corresponds to $r_{ij} = 1$. The inner boundary corresponds to $r_{ij} = -1$ if showing all associations, and to $r_{ij} = 0$ if showing positive associations only. Instead of the dot representation originally used, we suggest a frequency-based representation to visualize the row-column associations. A histogram H_j is created in each sector j to show the distribution of the associations r_{ij} along the radial dimension. An adjustable number b of equal bins is used for all histograms, initially determined by Scott's normal reference rule [131]. Each bin k in sector j aggregates the rows i having associations in the interval $I_k = [l_k, l_{k+1}[$. The interval boundaries l_k are equally spaced between $[-1, 1]$:

$$l_k = \frac{2(k-1) - b}{b} \tag{C.7}$$

A closed interval $I_b = [l_b, 1]$ is used for the last bin to account for $r_{ij} = 1$. Hence, the number of items h_{kj} in the k^{th} bin of sector j is:

$$h_{kj} = |\{1 \leq i \leq n : f_{i+} \geq T_s \wedge r_{ij} \in I_k\}| \tag{C.8}$$

Each bin k of histogram H_j is visualized as a track in sector j. This track occupies the radial position which corresponds to its interval I_k. The length of this track is proportional to h_{kj}. A uniform or sector-specific scaling factor ensures that all tracks fit in their sectors. Tracks

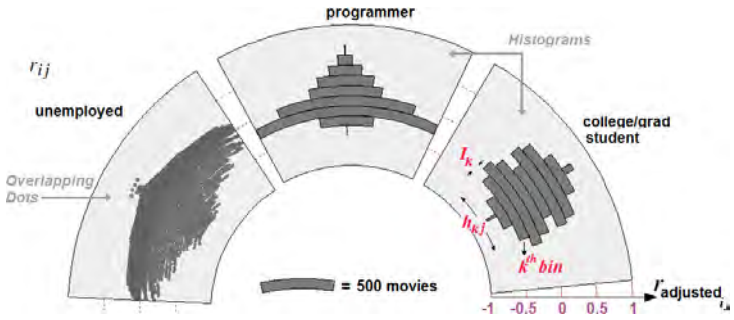


Figure C.4: Dot vs. histogram representation of row-column associations. The dimensions of a histogram bin are annotated (Eq. C.8).

are centered in their sectors, following the Gestalt principle of symmetry [165]. This avoids artificial asymmetry along the angular dimension in the sectors and makes it easier to compare their histograms. Figure C.4 shows both dot and histogram representations for some sectors of figure C.3c. The histograms show how 3706 movies are associated with 2 occupations. Both positive and negative associations are included.

Rows whose associations with sector j lie in a specific interval can be inspected individually along with the attributes of their entities, as explained in Sect. C.2.3. The distribution of a numerical or categorical attribute of these entities can be shown by coloring the histograms instead of coloring individual dots. This provides a clearer understanding of the attribute distribution at different radial distances. Figure C.5a shows the release-date distribution of movies positively associated with specific occupation categories. Figure C.5b shows the genres of the movies. Movies highly associated with the "Retired" category tend to be old. The opposite holds for the "K-12 student" category which also tends to be highly associated with "Children" movies. Movies highly associated with "Technician/Engineer" are more likely to have "Sci-Fi / Fantasy" genres. These tendencies seem stronger as compared to the distribution of both attributes among all movies (figure C.5c).

The frequency-based representation has several advantages over the dot representation: First, the angular dimension now has a clear meaning (frequency of associations at different radial distances in the sectors). Second, the artifacts and overlaps caused by showing separate dots are eliminated. Third, histograms are familiar visualizations that are easy to interpret. They better emphasize that the visualization is showing a distribution of the row associations in each sector, and not individual entities. This avoids the confusion due to multiple dots representing the same row. Finally, the redundancy of double-coding the association using both dot size and dot location is also eliminated.

Bended histograms embedded in a ring chart suffer from visual illusions in perceiving different arc lengths at different radial distances. This effect can be accounted for computationally and is minimized when arcs are short that are perceptually flattened [123] (figure C.6).

Visualizing column similarities

We compute similarities between the columns of the contingency table based on their row associations. A similarity value $rc_{j_1 j_2}$ is computed for every pair of columns (j_1, j_2), to assess how similar the two distributions r_{ij_1} and r_{ij_2} are. Only active rows in both sectors are included in the computation. Active rows in sector j have sufficient support f_{i+} and associations r_{ij} higher than T_r, and are defined as follows:

$$A_j = \{1 \leq i \leq n : r_{ij} \geq T_r \wedge f_{i+} \geq T_s\} \tag{C.9}$$

Active rows in each sector are depicted in dark gray in the respective histogram (figure C.6b). The column similarities are computed as follows:

$$rc_{j_1 j_2} = \frac{1}{|A_{j_1}| + |A_{j_2}|} \cdot \sum_{i \in A_{j_1} \cap A_{j_2}} r_{ij_1} \cdot r_{ij_2} \tag{C.10}$$

Between each pair of sectors (j_1, j_2), an arc is drawn whose thickness and opacity are determined by $rc_{j_1 j_2}$. A thick arc means that the active rows in both sectors tend to have similar associations with the two columns j_1 and j_2. Changing the T_r value results in smaller or larger active parts, and hence influences the thicknesses. By checking the arcs with different T_r values, the user can examine in which ranges and to which extent the column similarities hold.

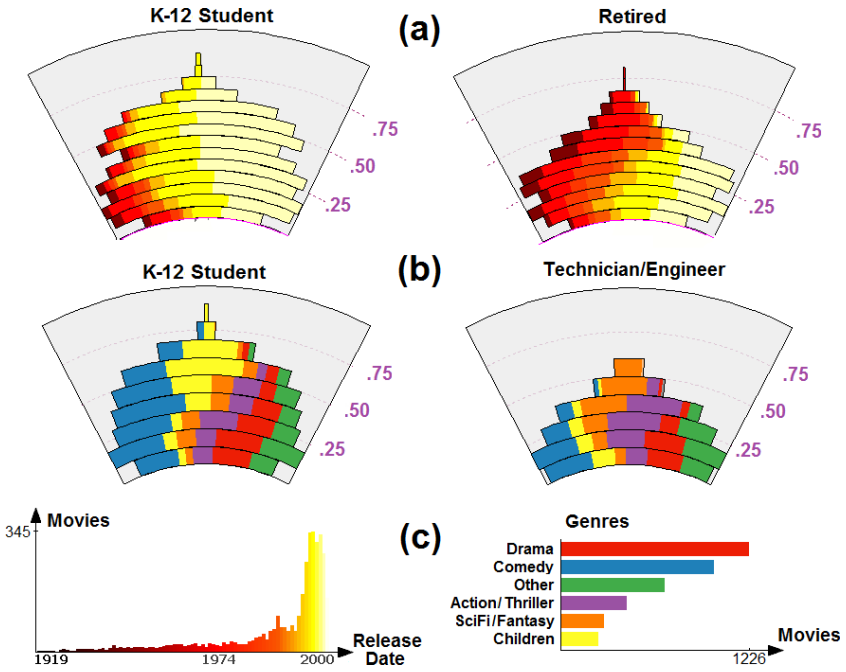


Figure C.5: Distributions of (a) a numerical attribute (release date) or, (b) a categorical attribute (genre) of the movies in the histograms. (c) The global distributions of release date and genre among all movies.

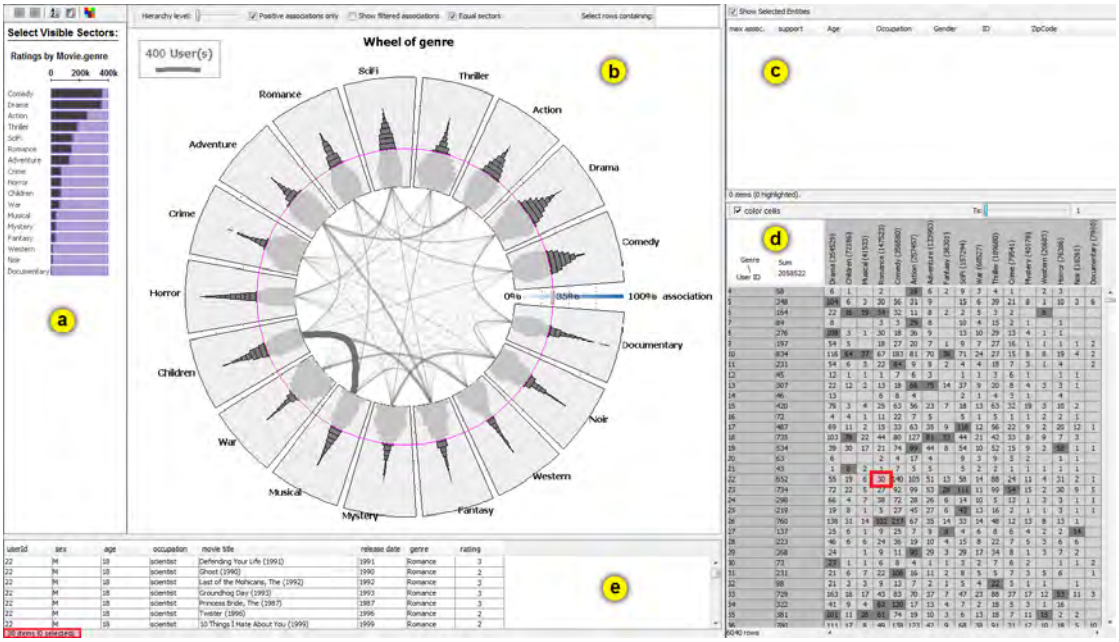


Figure C.6: Five levels of abstraction to explore the user-genre table and the underlying information: (a) a bar chart of the column categories (genres), (b) the wheel view showing sectors for the items selected in (a), (c) detail view for items selected in (b) (currently empty), (d) the contingency table with cells in active parts in (a) colored in dark gray, (e) the categorical data summarized in the cell highlighted in red in (d).

Arcs showing column similarities based on row associations is a unique feature of the Contingency Wheel and one of the main reasons of adopting a circular layout for the visualization. This layout provides a compact representation to show and compare column similarities. Furthermore, arcs are useful in creating a user-controlled hierarchical grouping of the column categories based on their similarities: A right-click on an arc merges the two affected sectors into one sector. The resulting wheel is built from the contingency table that results by merging the corresponding columns into one column, by summing up the frequencies cell-by-cell. The new sector is inserted at its appropriate position according to the sector ordering scheme currently in use (alphabetical, by size, or user-defined sector ordering). Successively merging pairs of sectors connected by thick arcs enables abstracting the visualization by reducing the number of visual items. Moreover, it enables analyzing similarities between groups of similar columns and not only between pairs of columns, as the use case shows (Sect. C.3).

Visual aids

We provide several visual aids to facilitate understanding. Three association levels evenly spaced between the inner and the outer sector boundaries are shown to allow an easier interpretation of

the radial distances. An additional circle in pink shows the current value of the association threshold T_r, which can be adjusted using the slider embedded in the ring chart. Inactive parts of the histograms (Eq. C.9) are visually de-emphasized. A color gradient is shown in the background of the T_r slider to reflect the association levels. It uses either a diverging or a sequential color scale [57], depending on whether negative associations are included or not. Arcs outside the ring chart indicate sector groups (figure C.1). Finally, a legend shows the scale used in the histograms by depicting an arc of average length.

C.2.3 Interactive Exploration Environment

The original Contingency Wheel may result in a cluttered visualization especially for large data because it creates dots for individual row entities. These dots need to be selected individually to obtain details about the corresponding entities [6]. To improve on these shortcomings, our new methods follow Shneiderman's visual information-seeking mantra [137]: The visualization first shows an overview of the data using histograms. The user can filter the data interactively and select entities she is interested in exploring. Then, details about these entities can be obtained in linked views. Contingency Wheel++ offers an overview visualization of an asymmetrically-sized contingency table. Likewise, the contingency table offers a summarization of a larger data set by cross-tabulating two categorical dimensions. We designed the user interface to enable exploring the data at these multiple levels of abstraction as explained in the following.

Multiple Views

Whenever we explain Contingency Wheel++ to new users, our first step is to show the underlying contingency table. This allows explaining the basic concepts like row- and column marginal frequencies, actual- and expected frequencies (Eq. C.1), and row-column associations (Eq. C.2-C.5). We are thus showing both the wheel visualization and the underlying table side-by-side in one interface. This combination bridges the gap between the visual representations and the data representation (i.e., association values) computed by the automatic methods. The main user interface (UI) of our prototype is divided into five coordinated views:

A *bar chart* shows the column categories and their marginal frequencies f_{+j} (figure C.6a). Columns selected in this view define the sectors of the wheel view. The user can thus focus on selected columns. Also, if the number of columns exceeds the limits for the wheel, smaller subsets of columns can still be visualized.

The *wheel view* is the central part of the interface (figure C.6b). It provides an overview of the data and existing associations within. Several interactions are possible to find interesting patterns in the data and select specific row entities for further analysis. The association threshold T_r can be adjusted interactively via the slider embedded in the ring chart. Also, this view enables setting several parameters by means of its toolbar and context menu.

A *list view* shows details about the row entities selected in the wheel view (figure C.6c and figure C.7d). Beside the attributes of these entities (available from the data set), their marginal row frequencies f_i+ and associations r_{ij} with a specific column j are listed. The entities can be sorted according to one of the columns, and histograms or bar charts can be created for a specific column in the list.

A *tabular view* shows the contingency table and the marginal frequencies (figure C.6d). By hovering the mouse pointer over a cell (i, j), a tooltip shows the expected frequency $\hat{e}_{ij}$ and the association value r_{ij} according to the measure used. If cell coloring is enabled via a checkbox, the cell is shown in dark gray if it corresponds to an active part in the visualization (i.e., $i \in A_j$). Also, the support threshold T_s can be adjusted via a slider to filter out entire rows i with $f_{i+} < T_s$.

A *second list view* shows details about selected items from the tabular view (figure C.6e). By double-clicking on a cell, a row, or a column in the tabular view, cross-tabulated data items are shown in this list view along with their attributes. The items can be sorted and the distributions of the values in a specific column can be explored using a histogram or a bar chart.

These views make it easier to explain to new users how the data are visualized in Contingency Wheel++. Even more importantly, they constitute a multi-level overview+detail exploration interface. This allows experienced users to perform elaborate analysis workflows by having quick access to all information available in the data. Hence, associations can be detected and investigated further in relation to other attributes. The incorporation of analytical methods in the visual interface enables a visual analytics process following Keim's mantra [79]: Analyze first – show the important – zoom, filter and analyze further – details on demand. After computing the row-column associations (Sect. C.2.1) and the columns similarities (Sect. C.2.2), the visualization shows the important results, i.e., strong associations or high similarities. Using different interactions, the user can change the thresholds T_r and T_s, merge columns, or set a different association measure. This causes the analytical methods to recompute the associations and similarities which are then visualized interactively. Details on selected items in the wheel or in the tabular view can be obtained on demand.

Linking and Brushing

Contingency Wheel++ offers multiple ways to brush the visualized row categories. One way is by clicking on a bar in the histograms, which selects the rows it aggregates (Sect. C.2.2). Another way is using the sector marquee tool to define a radial interval I in a sector j using the mouse (figure C.1). This selects the rows i with $r_{ij} \in I$. Clicking on sector j selects the rows A_j that are currently active (Sect. C.2.2). Also, clicking on an arc selects rows active in both sectors it connects (the items that define this arc). Rows i with $r_{ij} \leq T_r$ for all columns j can be selected using a menu command. When T_r is positive but small, this command selects rows that do not exhibit a high association with any column. Finally, rows can be selected using an external query, like the instant search box at the top of the view. This box selects row categories containing a specific text.

The top-right list view (figure C.7d) shows the selected rows defined either by filtering, brushing, selection, or the search box query. When an item in this list is clicked, the tabular view scrolls to and highlights the corresponding row i which shows the frequencies f_{ij} (figure C.12e). Also, a star graph [56] of the associations r_{ij} can be shown in the wheel view, labeled with these frequencies. Selected row categories are highlighted in the histograms of all sectors. The original histograms become desaturated and new sub-histograms are drawn centered on top of them showing the selected portion using color (figure C.7c). Likewise, the original arcs are

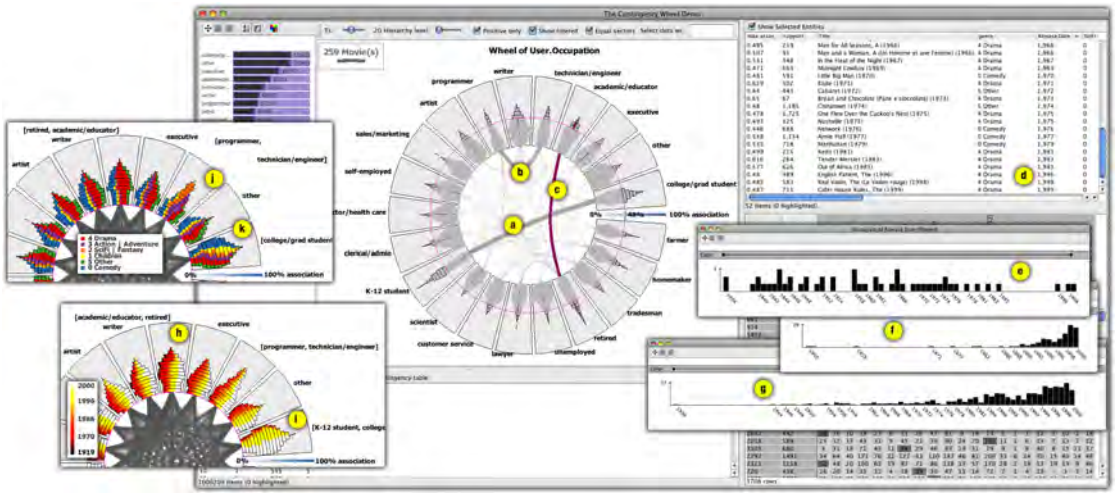


Figure C.7: Visual exploration of movies associated to user occupations: (a-c) major overlaps between user groups, (d) details of selected items in (c), (e-g) histograms of movie release date for different subsets of movies, (h, i) wheel view colored by movie release dates to reveal its relation with different user groups, (j, k) wheel view colored by movie genre to reveal dominant genres in the movie preferences of different user groups.

desaturated and the parts corresponding to the selected items are highlighted. Three modes are offered for performing brushing operations in the wheel, depending on keyboard modifiers:

- Set union: the new selection is added to an existing selection.
- Set intersection: the new selection is intersected with the existing selection. This enables creating nested queries on the data. For example, in the wheel showing the movies-occupation table, the user can select movies highly associated with the categories “programmer” and “scientist” but negatively associated with the category “executive”. This is done by drawing ranges at the corresponding radial distances in each sector while the CTRL key is pressed. TimeSearcher uses a similar brushing technique for time-series data by means of timebox widgets [66].
- If no modifier is defined or if brushing is performed using an external query, the active selection is replaced by the new one.

C.3 Use Case

To demonstrate the applicability of our approach, we present a use case along the fictitious character Jane, who is an analyst at a large movie rental service. The use case is based on 10 analysis sessions conducted over the course of a week and added up to a total of 8 hours time. For the analysis, the MovieLens data set [53] has been used as introduced in Sect. 1. Jane’s goal is to get insights into the massive amount of data they have collected about their customers

who rented, watched, and rated movies using their service. Based on the insights gained from this analysis she plans to make decisions and take actions related to the ongoing marketing strategies and recommender algorithms they have in place. Jane uses mainly two different tables for her analysis: first, *occupations* (movies $\times$ user occupations) and second, *genres* (users $\times$ movie genres). By exploring the interactive wheels and the associated views and diagrams, Jane gains a number of insights, some of which were expected but also some surprising ones. Before Jane starts the analysis, she asks herself about the semantics of the data – what do associations between the entities user and movie actually mean? A user and a movie are associated if a user has entered a rating for a movie in the system which in turn implies that he or she has watched the movie. However, an association does not express how much they liked a movie.

C.3.1 Categories & Characteristics

As a first step, Jane aims for getting an overview of the categories to get a feeling for the data and overall (dis)similarities, to explore characteristics of single categories as well as to possibly simplify the data by merging categories.

User occupations Jane starts her analysis by creating a wheel based on a contingency table that displays the different occupations (i.e., jobs) of the users as sectors and the related movies as histogram bars within these sectors (figure C.7). After opening the initial wheel, she adjusts the association slider to $>$40% in order to focus on higher associations and similarities between sectors. By studying the thickness of the connections inside the wheel she notices that there is a lot of overlap (same movies rated) between "K-12 student" and "college/grad student " (figure C.7a) as well as between "programmer" and "technician/engineer" (figure C.7b) which seems plausible to her. However, a more surprising aspect is that there is also a higher degree of overlap between "academic/educator" and "retired" (figure C.7c). To investigate this connection in more detail, Jane selects the arc (figure C.7c) and takes a look at the selected entities in the list view in the upper right of the UI (figure C.7d). She sorts the movies by release date by clicking on the respective table header and discovers that they seem to be mostly older movies. Only three out of 52 movies are from the 90's, the rest are older movies. Based on the mentioned similarities, she merges "K-12 student" with "college/grad student", "programmer" with "technician/engineer", and "academic/educator" with "retired" by right-clicking on the corresponding arcs in order to simplify the further analysis.

Then, Jane continues her exploration of release dates of movies highly associated to the group [academic/educator, retired]. For this, she brings up a histogram using the context menu of the release-date column-header (figure C.7e). This provides details about the distribution of movies over time. For comparison, she also brings up release-date histograms for [college/grad student, K-12 student] (figure C.7f), as well as for all movies that were rated (figure C.7g). This confirms that the distribution concerning [academic/educator, retired] is quite different. Moreover, Jane finds out that there seems to be a peak of watched and rated movies in the mid 80s followed by a valley at the beginning of the 90s and another peak at the end of the 90s. To get a further overview of release dates of categories, she colors the wheel by release date which clearly shows that [academic/educator, retired] more often watch older movies than

others (larger portion of dark parts than average, figure C.7h). [K-12 student, college/grad student] watch more often newer movies than others (larger portion of bright parts than average, figure C.7i). Jane concludes her exploration by coloring the occupation wheel by genre. This reveals that [programmer, technician/engineer] contain a much larger portion of highly associated "SciFi/Fantasy" movies (orange, figure C.7j) and [college/grad student, K-12 student] have a much larger portion of highly associated "Children" and "Comedy" movies (blue and yellow, figure C.7k) than on average.

Movie genres Jane switches to the genre wheel and finds high overlaps of "Musical" and "Children", "Action" and "Adventure", and "War" and "Western" which seem to be reasonable to her. More surprisingly, she finds no particularly high overlap between [War, Western] and "Crime" which she would have suspected. Besides, she observes that "Horror" seems inversely related to many other genres. Based on her observations she merges genres into [Children, Musical, and Fantasy], [Noir, Mystery, Thriller], [Adventure, Action, SciFi], and [War, Western] (figure C.3.1a-b). Jane colors the wheel by age using a diverging color scheme (figure C.3.1a). Looking at this, she finds it surprising that the age distribution of users watching [Musical, Children, Fantasy] is not very different from others. Overall, age-group distributions seem to be quite similar in all genres.

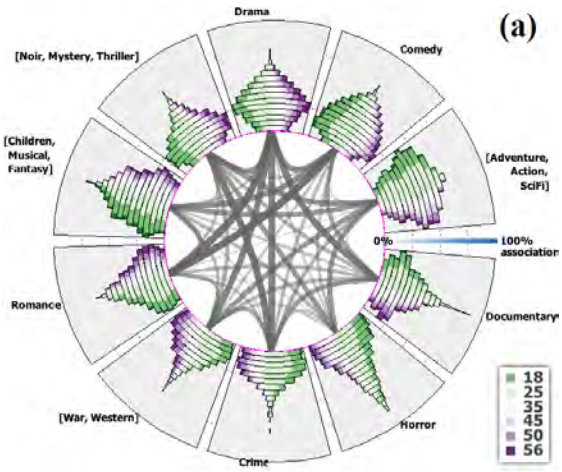


Figure C.8: Associations between users and movie genres: colored by age.

Further, Jane wants to inspect possible gender differences and turns on coloring by gender in the genre wheel (figure C.3.1b). As a general observation, she recognizes that there are a lot more men than women rating movies. As anticipated, Jane finds the genre "Romance" as an exception where the most highly associated users are female. Surprisingly "Horror" does not show less women than other genres such as "Documentary", [Adventure, Action, SciFi], [Noir, Mystery, Thriller], or "Crime". After that, Jane takes a closer look on different genres using histograms of gender, age, and occupations (figure C.3.1c). For "Children" movies she notices that there is an almost equal distribution between male and female viewers, and that most viewers are in the age group of 18–24 years. Particularly, the last fact is somewhat surprising to Jane, since she

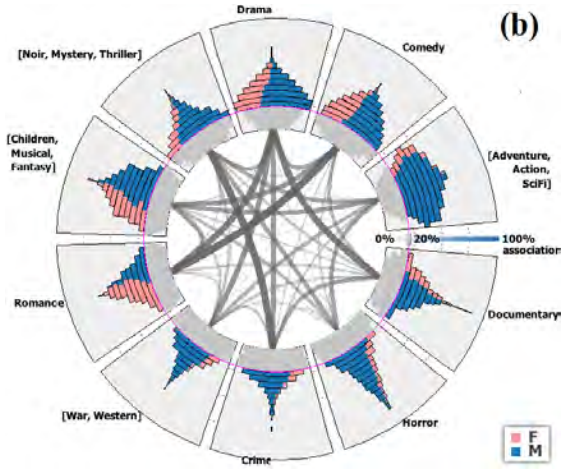


Figure C.9: Associations between users and movie genres: colored by gender.

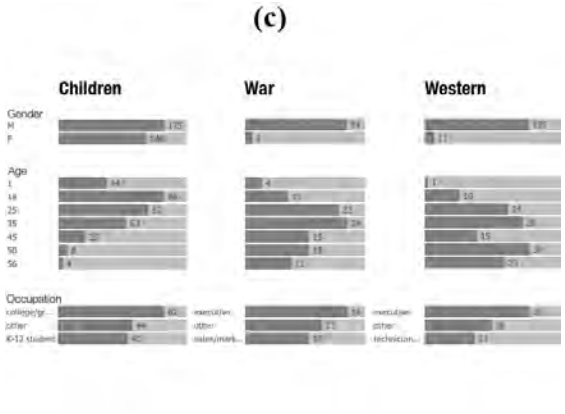


Figure C.10: Associations between users and movie genres: Details about selected genres.

thought that the majority of users watching "Children" films would be younger. Having a look at "War" movies she spots that there are by far more men present which are often executives and in the age group of 35–44 years. "Western" movies show a quite similar picture, except that even older age groups watch and rate these movies.

C.3.2 Single Movies

After her top-down exploration of occupations and genre categories, Jane has a couple of movies in mind she wants to inspect further for potentially interesting findings in a bottom-up manner.

Mainstream erotic films She takes a look at the two movies Basic Instinct (1992) and Nine 1/2 Weeks (1986) and compares the star plots in the wheel views. Interestingly, both movies

are highly associated to “technician/engineer” but negatively associated to “programmer” (figure C.11a-d) which are otherwise quite similar as she had found out earlier.

The Godfather trilogy Next, Jane remembers that The Godfather movies (1972, 1974, 1990) were quite big hits at her movie rental service in the last years. She uses the search box (figure C.12a) to find them. The movies matching the query are shown in the detail list in the upper right of the UI. She selects the first movie of the trilogy in the list (figure C.12b) which brings up a star plot in the center of the wheel view showing individual associations for the different occupations. She can see that the movie has the highest associations with the occupations “executive” and “lawyer” (figure C.12c,d). When she selects the second movie, the picture is quite similar, however, the third movie is somewhat different. Jane sees that it is highly associated to “executive” again but that it is negatively associated to “lawyer” and highly associated to “sales/marketing”. Further, Jane would like to inspect how the three movies were rated among executives. For this, she double clicks on the “executive” column in the table (figure C.12e) which brings up the ratings in the list view on the lower left of the UI (figure C.12f). Via a context menu, she displays the rating histograms (figure C.12g-i) and spots that they are quite positive and similar for the first two but much lower for the third movie.

C.3.3 Hypotheses and Specific Questions

During the visual exploration, Jane generated some hypotheses and specific questions she is trying to answer subsequently.

Association vs. rating behavior One hypothesis Jane had in mind is to check whether it is true that very high associations of movies correspond to more positive ratings. Using the occupation wheel, she is probing rating histograms of highly associated movies (>75%) with “college/grad student”, such as Transformers (1986, good ratings) and Teenage Mutant Turtles II (1991, bad ratings). As a result, she finds evidence that her hypothesis does not hold.

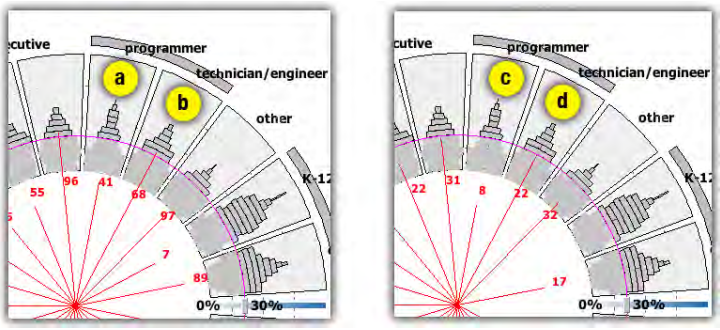


Figure C.11: Associations of mainstream erotic films to “programmer” (a, c) and “technician/engineer” (b, d) – left: Basic Instinct (1992), right: Nine 1/2 Weeks (1986).

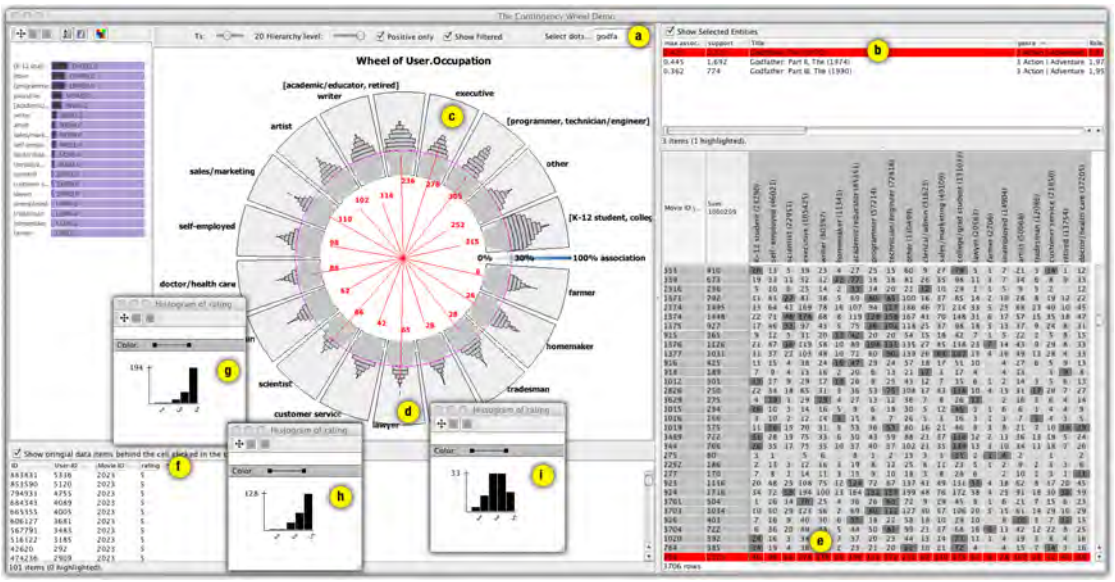


Figure C.12: Visual exploration of the Godfather trilogy: (a) search box, (b) search result, (c, d) star plot of associations of selected item to user occupations, (e) row of selected movie in contingency table, (f) raw data of user ratings, (g-i) rating histograms for the three movies of the trilogy.

Rating Another question Jane wants to inspect is whether there are differences in the general rating behavior of different user occupations, i.e., are particular groups more or less critical than others in general? By using the occupation wheel and comparing the grading histograms of selected sectors, Jane observes that the rating behavior is strikingly similar among groups. She can only spot subtle differences such as that unemployed persons tend to give lower ratings whereas retired persons do not tend to give many low ratings.

C.3.4 Decisions and Actions Planned

Visually exploring the vast data collection of her movie rental service helped Jane to better understand her customers and unearth commonalities as well as differences between groups of users and movies. Based on the gained insights, decisions are taken and actions are planned that are intended to make her business more successful: The merging of some user categories and movie genres can simplify the internal recommender engine. New SciFi & Fantasy releases will be presented particularly to programmers and technicians/engineers. As there were more movies watched and rated from the mid 1980s, there will be a campaign highlighting some of these. Suggestions concerning Children movies will no longer be focused on younger customers but concentrate on the age group of 18–24 years. War and Western movies will be recommended more intensively to male executives older than 35 years. Finally, Horror movies will be suggested to users who already watched those without restricting suggestions to men.

C.3.5 Improvements of Contingency Wheel++

Jane benefited from the improvements in the new design and gained insights that would not have been possible using the original Contingency Wheel. Due to the fact that dots have been replaced by histograms, she was able to represent all movies without filtering steps which would have been necessary to avoid overlaps. The distribution of an attribute of the movies (e.g., release date) can now be inspected using colored histograms which allowed for complex insights involving multiple data attributes. Because of the multi-level overview+detail exploration environment, Jane had easy on-demand access to all available data, such as movie details, contingency table, and raw data. This allowed for drilling down to clarify and check findings from an aggregate level. Further, she was able to create bar charts and histograms of selected elements from different attributes, such as movie release dates or ratings, and compare the results with the global distributions. On top of that, the ability to merge sectors enabled the detection of patterns between groups of sectors that could not be detected when looking at single columns.

C.4 State of the Art

Methods dedicated for visualizing contingency tables are usually designed to handle a small number of categories. Based on what the visual representations depict, they can be classified into three types:

Frequency representations These methods map the table frequencies f_{ij} to visual elements of proportional size. Mosaic Displays [58] and their variations use tiles to represent the frequencies (similar to Treemaps [138]). Parallel Sets [13] and their variations, such as Circos [87], represent frequencies as stripes or ribbons between visual elements that depict the categories. These approaches offer an intuitive visual representation that can be divided further to accommodate additional dimensions. However, they can handle only a relatively small number of categories (≤ 30 for Parallel Sets). With larger tables, the clutter increases in Parallel Sets, and the increased skewness and number of zeros in the table values make it difficult to identify and compare the tiles in Mosaic Displays [156].

Deviation representations Association Plots [100] use bar charts to show the deviations between the actual frequencies f_{ij} and the expected frequencies $\hat{e}_{ij}$ (Eq. C.1). Sieve Diagrams [49] plot f_{ij} as sieves in Mosaic Displays of $\hat{e}_{ij}$ to show how both deviate from each other. Sieves with smaller holes represent higher associations. A recent approach was proposed for exploring proportions in multivariate categorical data [112]. It adopts the layout of Parallel Sets, but depicts the proportionality of relationships between the categories instead of f_{ij}.

Intermediate representations Correspondence Analysis [14] (CA) projects the categories to points in a 2D space, spanned by the two most contributing factors of the χ^2 statistic, in a way similar to Principal Component Analysis [75]. A higher association between categories of the same class positions their points closer together, in a way similar to multidimensional scaling [86]. The approach can also accommodate additional categorical dimensions [52]. However,

with a growing number of categories, the plot becomes more difficult to read. It lacks an intuitive structure as its axes bear no interpretable semantics. Johansson et al. [74] and Rosario et al. [126] proposed methods for quantifying categorical data based on CA. The quantified data can then be visualized using scatter plots or parallel coordinates. However, the latent numerical variables used for the quantification are not always easy to interpret.

The dot-based Contingency Wheel uses deviation representations for the cells as dots along the radial dimension. Like many approaches for dealing with large data [37, 117, 145] it uses data reduction to handle tables having a large number of rows. Also, it employs alpha blending to reveal overlapping, as done by other approaches for dealing with similar issues [44, 73, 85]. In contrast, Contingency Wheel++ employs a frequency-based approach to abstract large data, as used by many other techniques [60, 84, 125]. Also, it makes use of interactive visual analytics techniques to enable the exploration of individual data items. As the use case illustrates, asymmetrically-sized tables with a small number of columns (≤ 30) and thousands of rows can be handled efficiently by Contingency Wheel++ without filtering the data.

C.5 Conclusion

Contingency Wheel++ employs novel visual analytics methods that address the major shortcomings of the original dot-based wheel for visualizing and discovering patterns in large categorical data. It improves on the computational component by introducing an association measure based on Pearson's residuals to alleviate the bias in the association measure originally used. It eliminates the scalability and readability limitations caused by overlapping dots, by using a frequency-based abstraction that shows distributions rather than individual entities. Finally, it offers a multi-level overview+detail interface to explore individual entities that are aggregated in the visualization or in the table along with their attributes. The use case demonstrates how these methods can be used to find nontrivial patterns in large categorical data, and how further attributes can be analyzed in separate views or by coloring the histograms in the wheel visualization.

Future work aims to conduct comparative user studies to assess the effectiveness and efficiency of Contingency Wheel++, and to apply it to different real-world domains. Also, we are exploring further measures of associations and column similarities. Finally, we are investigating the applicability of our approach to other problems having similar data structures, such as point-set memberships or the class probabilities computed by a fuzzy classifier for a large number of samples.

Bibliography

[1] The ACM computing classification system [1998 version]. `http://www.acm.org/about/class/1998`. accessed: August 2014.

[2] The IMDB database (snapshot in sept. 2012). `http://www.imdb.com/interfaces`. accessed: August 2014.

[3] B. Alper, N. Henry Riche, G. Ramos, and M. Czerwinski. Design study of linesets, a novel set visualization technique. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2259–2267, 2011.

[4] B. Alsallakh, W. Aigner, S. Miksch, and Meister Eduard Gröller. Reinventing the Contingency Wheel: Scalable visual analytics of large categorical data. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2849–2858, 2012.

[5] Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and Helwig Hauser. Radial Sets: Interactive visual analysis of large overlapping sets. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2496–2505, 2013.

[6] Bilal Alsallakh, Eduard Gröller, Silvia Miksch, and Martin Suntinger. Contingency wheel: Visual analysis of large contingency tables. In *Proceedings of the International Workshop on Visual Analytics (EuroVA)*, pages 53–56. Eurographics, 2011.

[7] Bilal Alsallakh, Allan Hanbury, Helwig Hauser, Silvia Miksch, and Andreas Rauber. Visual analytics methods for probabilistic classification data. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12), 2014. to appear.

[8] Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch, and Peter Rodgers. Visualizing sets and set-typed data: State-of-the-art and future challenges. In *Eurographics conference on Visualization (EuroVis)?State of The Art Reports*. Eurographics, 2014.

[9] Bilal Alsallakh, Silvia Miksch, and Andreas Rauber. Towards a visualization of multi-faceted search results. In *Proceedings of the DL2014 Workshop on Knowledge Maps and Information Retrieval (KMIR), the ACM/IEEE Joint Conference on Digital Libraries*, 2014.

[10] A. Anand, L. Wilkinson, and Dang Nhon Tuan. An L-infinity norm visual classifier. In *IEEE International Conference on Data Mining (ICDM)*, pages 687–692, 2009.

[11] Mihael Ankerst, Christian Elsen, Martin Ester, and Hans-Peter Kriegel. Visual classification: an interactive approach to decision tree construction. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 392–396. ACM, 1999.

[12] K. Bache and M. Lichman. UCI machine learning repository, 2013. University of California, Irvine, School of Information and Computer Sciences. `http://archive.ics.uci.edu/ml`.

[13] Fabian Bendix, Robert Kosara, and Helwig Hauser. Parallel sets: visual analysis of categorical data. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 133–140, 2005.

[14] J. P. Benzécri. *Correspondence Analysis Handbook*. Marcel Dekker, New York, 1990.

[15] Jürgen Bernard, Martin Steiger, Sven Widmer, Hendrik Lücke-Tieke, Thorsten May, and Jörn Kohlhammer. Visual-interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets. *Computer Graphics Forum*, 33(3):291–300, 2014.

[16] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz information miner. In *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 319–326. Springer Berlin Heidelberg, 2008.

[17] Jacques Bertin. *Graphics and graphic information processing*. de Gruyter, 1981.

[18] Jacques Bertin and Myriam Daru. Matrix theory of graphics: Jacques Bertin's theories. *Information Design Journal*, 10(1):5–19, 2000.

[19] Nadia Boukhelifa and Peter J Rodgers. A model and software system for coordinated and multiple views in exploratory visualization. *Information Visualization*, 2(4):258–269, 2003.

[20] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2376–2385, 2013.

[21] Sebastian Bremm, Tatiana von Landesberger, Jürgen Bernard, and Tobias Schreck. Assisted descriptor selection based on visual comparative data analysis. *Computer Graphics Forum*, 30(3):891–900, 2011.

[22] E.T. Brown, Jingjing Liu, C.E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 83–92, 2012.

[23] Doina Caragea, Dianne Cook, Hadley Wickham, and Vasant Honavar. Visual methods for examining SVM classifiers. In *Visual Data Mining*, pages 136–153. Springer, 2008.

[24] Stuart K Card, Jock D Mackinlay, and Ben Schneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[25] Remco Chang, Mohammad Ghoniem, Robert Kosara, William Ribarsky, Jing Yang, Evan Suma, Caroline Ziemkiewicz, Daniel Kern, and Agus Sudjianto. WireVis: Visualization of categorical, time-varying data from financial transactions. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 155–162. IEEE, 2007.

[26] Chun-Houh Chen. Generalized association plots: Information visualization via iteratively generated correlation matrices. *Statistica Sinica*, 12(1):7–30, 2002.

[27] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *IEEE Symp. on Visual Analytics Science and Technology (VAST)*, pages 27–34, 2010.

[28] Stirling Christopher Chow. Generating and drawing area-proportional Euler and Venn diagrams. *PhD Thesis at the University of Victoria - Canada*, 2007.

[29] W.S. Cleveland. *The elements of graphing data*. AT&T Bell Laboratories, 1994.

[30] Andy Cockburn, Amy Karlson, and Benjamin B Bederson. A review of overview+ detail, zooming, and focus+ context interfaces. *ACM Computing Surveys (CSUR)*, 41(1):2, 2008.

[31] C. Collins, G. Penn, and S. Carpendale. Bubble sets: Revealing set relations with isocontours over existing visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1009–1016, 2009.

[32] Isabel F Cruz and Yuan Feng Huang. A layered architecture for the exploration of heterogeneous information using coordinated views. In *Visual Languages and Human Centric Computing, IEEE Symposium on*, pages 11–18. IEEE, 2004.

[33] Amy E Daniels. Incorporating domain knowledge and spatial relationships into land cover classifications: a rule-based approach. *International Journal of Remote Sensing*, 27(14):2949–2975, 2006.

[34] Jay Devore. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 2011.

[35] Emilio Di Giacomo, Luca Grilli, and Giuseppe Liotta. Drawing bipartite graphs on two curves. In *Graph Drawing*, pages 380–385. Springer, 2007.

[36] K. Dinkla, M.J. van Kreveld, B. Speckmann, and M.A. Westenberg. Kelp diagrams: Point set membership visualization. In *Computer Graphics Forum*, volume 31, pages 875–884. Wiley Online Library, 2012.

[37] Alan Dix and Geoff Ellis. By chance: enhancing interaction with large data sets through statistical sampling. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '02, pages 167–176, New York, NY, USA, 2002. ACM.

[38] Helmut Doleisch. SimVis: Interactive visual analysis of large and time-dependent 3d simulation data. In *Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come*, pages 712–720. IEEE Press, 2007.

[39] Chris Drummond and RobertC. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

[40] Peter Eades and Nicholas C Wormald. Edge crossings in drawings of bipartite graphs. *Algorithmica*, 11(4):379–403, 1994.

[41] Niklas Elmqvist, Thanh-Nghi Do, Howard Goodell, Nathalie Henry Riche, and J-D Fekete. ZAME: Interactive large-scale graph visualization. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 215–222. IEEE, 2008.

[42] Leonhard Euler. *Lettres à une princesse d'Allemagne sur divers sujets de physique et de philosophie*, volume 1 letters no. 102-108. Courcier, 1772.

[43] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[44] J.-D. Fekete and C. Plaisant. Interactive information visualization of a million items. In *IEEE Symposium on Information Visualization, 2002.*, pages 117–124, 2002.

[45] Stephen Few. Our irresistible fascination with all things circular. *Perceptual Edge Visual Business Intelligence Newsletter*, pages 1–9, 2010.

[46] Jean Flower, Andrew Fish, and John Howse. Euler diagram generation. *Journal of Visual Languages & Computing*, 19(6):675–694, 2008.

[47] Jean Flower and John Howse. Generating Euler diagrams. *Diagrammatic Representation and Inference*, pages 285–285, 2002.

[48] Wolfgang Freiler, Kresimir Matkovic, and Helwig Hauser. Interactive visual analysis of set-typed data. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1340 – 1347, November 2008.

[49] M. Friendly. Graphical methods for categorical data. In *SAS User Group International Conference Proceedings*, volume 17, pages 190–200, 1992.

[50] Karl Ruben Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[51] Mohammad Ghoniem, J-D Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization (INFOVIS)*, pages 17–24. IEEE, 2004.

[52] Michael J. Greenacre and Jörg Blasius. *Multiple correspondence analysis and related methods*. Chapman & Hall/CRC, 2006.

[53] GroupLens. MovieLens data sets. http://www.grouplens.org/node/73. accessed: August 2014.

[54] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[55] Lieve Hamers, Yves Hemeryck, Guido Herweyers, Marc Janssen, Hans Keters, Ronald Rousseau, and André Vanhoutte. Similarity measures in scientometric research: the jaccard index versus salton's cosine formula. *Information Processing & Management*, 25(3):315–318, 1989.

[56] Robert L. Harris. *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, Inc., New York, NY, USA, 1999.

[57] Mark Harrower and Cynthia Brewer. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, pages 27–37, June 2003.

[58] J. A Hartigan and B Kleiner. Mosaics for contingency tables. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 268–273. Springer-Verlag, 1981.

[59] Md Rafiul Hassan, Kotagiri Ramamohanarao, Chandan Karmakar, M Maruf Hossain, and James Bailey. A novel scalable multi-class ROC for effective visualization and computation. In *Advances in Knowledge Discovery and Data Mining*, pages 107–120. Springer, 2010.

[60] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *IEEE Symposium on Information Visualization*, pages 127 – 130, 2002.

[61] Helwig Hauser, Florian Ledermann, and Helmut Doleisch. Angular brushing of extended parallel coordinates. In *IEEE Symposium on Information Visualization (INFOVIS)*, pages 127–130. IEEE, 2002.

[62] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *Visualization and Computer Graphics, IEEE Trans. on*, 18(12):2839–2848, 2012.

[63] N Henry Riche and T. Dwyer. Untangling Euler diagrams. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1090–1099, 2010.

[64] Nathalie Henry Riche and J-D Fekete. MatrixExplorer: a dual-representation system to explore social networks. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):677–684, 2006.

[65] Nathalie Henry Riche and Jean-Daniel Fekete. MatLink: Enhanced matrix visualization for analyzing social networks. *Human-Computer Interaction–INTERACT 2007*, pages 288–302, 2007.

[66] Harry Hochheiser and Ben Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, March 2004.

[67] B. Hoferlin, R. Netzel, M. Hoferlin, D. Weiskopf, and G. Heidemann. Inter-active learning of ad-hoc classifiers for video visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 23–32, 2012.

[68] Patrick Hoffman, Georges Grinstein, Kenneth Marx, Ivo Grosse, and Eugene Stanley. Dna visual and analytic data mining. In *Visualization'97., Proceedings*, pages 437–441. IEEE, 1997.

[69] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[70] T. Itoh, C. Muelder, Kwan-Liu Ma, and J. Sese. A hybrid space-filling and force-directed layout method for visualizing multiple-category graphs. In *Visualization Symposium, 2009. PacificVis '09. IEEE Pacific*, pages 121–128, 2009.

[71] Tomoharu Iwata, Kazumi Saito, Naonori Ueda, Sean Stromsten, Thomas L Griffiths, and Joshua B Tenenbaum. Parametric embedding for class visualization. *Neural Computation*, 19(9):2536–2556, 2007.

[72] Mikael Jern, Sara Johansson, Jimmy Johansson, and Johan Franzen. The GAV toolkit for multiple linked views. In *Coordinated and Multiple Views in Exploratory Visualization (CMV)*, pages 85–97. IEEE, 2007.

[73] Jimmy Johansson, Patric Ljung, Mikael Jern, and Matthew Cooper. Revealing structure in visualizations of dense 2d and 3d parallel coordinates. *Information Visualization*, 5(2):125–136, June 2006.

[74] Sara Johansson, Mikael Jern, and Jimmy Johansson. Interactive quantification of categorical variables in mixed data sets. In *Proceedings of the 12th International Conference on Information Visualisation*, pages 3–10, Washington, DC, USA, 2008. IEEE Computer Society.

[75] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[76] Gaetano Kanizsa and Walter Gerbino. Convexity and symmetry in figure-ground organization. *Vision and artifact*, pages 25–32, 1976.

[77] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. Interactive optimization for steering machine classification. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, pages 1343–1352. ACM, 2010.

[78] Daniel Keim, Ming Hao, Umesh Dayal, Meichun Hsu, and Julain Ladisch. Pixel bar charts: A new technique for visualizing large multi-attribute data sets without aggregation. In *Information Visualization, IEEE Symposium on*, pages 113–113. IEEE Computer Society, 2001.

[79] Daniel Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual analytics: Scope and challenges. In *Visual Data Mining*, volume 4404 of *Lecture Notes in Computer Science*, pages 76–90. Springer Berlin / Heidelberg, 2008.

[80] Daniel A. Keim. Visual techniques for exploring databases. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) - Invited Tutorial*, 1997.

[81] Daniel A. Keim. Designing pixel-oriented visualization techniques: theory and applications. *Visualization and Computer Graphics, IEEE Transactions on*, 6(1):59–78, 2000.

[82] Hans Kestler, André Müller, Johann Kraus, Malte Buchholz, Thomas Gress, Hongfang Liu, David Kane, Barry Zeeberg, and John Weinstein. VennMaster: area-proportional Euler diagrams for functional GO analysis of microarrays. *BMC bioinformatics*, 9(1):67, 2008.

[83] Wolfgang Kienreich and Christin Seifert. Visual exploration of feature-class matrices for classification problems. In *International Workshop on Visual Analytics (EuroVA)*, pages 37–41. The Eurographics Association, 2012.

[84] Robert Kosara, Fabian Bendix, and Helwig Hauser. Timehistograms for large, time-dependent data. In Oliver Deussen, Charles Hansen, Daniel Keim, and Dietmar Saupe, editors, *Symposium on Visualization (VisSym)*, pages 45–54, 340. Eurographics Association, 2004.

[85] Robert Kosara, Silvia Miksch, and Helwig Hauser. Focus+context taken literally. *IEEE Computer Graphics and Applications*, 22:22–29, 2002.

[86] J B Kruskal and M Wish. Multidimensional scaling. *Methods*, 116(2):463–504, 1978.

[87] Martin Krzywinski, Jacqueline Schein, I.nanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009.

[88] Nicolas Lachiche and Peter Flach. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In *International Conference on Machine Learning (ICML)*, volume 20, pages 416–423, 2003.

[89] Louisa Lam and Ching Y Suen. Optimal combinations of pattern classifiers. *Pattern Recognition Letters*, 16(9):945–954, 1995.

[90] Jeffrey LeBlanc, Matthew O Ward, and Norman Wittels. Exploring n-dimensional databases. In *Proceedings of the 1st conference on Visualization'90*, pages 230–237. IEEE Computer Society Press, 1990.

[91] Haim Levkowitz. Color icons: merging color and texture perception for integrated visualization of multiple parameters. In *Proceedings of the 2nd conference on Visualization'91*, pages 164–170. IEEE Computer Society Press, 1991.

[92] Innar Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining*, 3(2):70–91, 2010.

[93] Zhicheng Liu, Shamkant B Navathe, and John T Stasko. Network-based visual analysis of tabular data. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 41–50. IEEE, 2011.

[94] Erkki Mäkinen. How to draw a hypergraph. *International Journal of Computer Mathematics*, 34(3-4):177–185, 1990.

[95] Erkki Mäkinen and Harri Siirtola. Reordering the reorderable matrix as an algorithmic problem. *Theory and Application of Diagrams*, pages 453–468, 2000.

[96] Frank J Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[97] Kresimir Matkovic, Wolfgang Freiler, Denis Gracanin, and Helwig Hauser. Comvis: A coordinated multiple views system for prototyping new visualization technology. In *Information Visualisation, 2008. IV'08. 12th International Conference*, pages 215–220. IEEE, 2008.

[98] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 111–120, 2011.

[99] W. Meulemans, N. Henry Riche, B. Speckmann, B. Alper, and T. Dwyer. KelpFusion: a hybrid set visualization technique, 2013.

[100] David Meyer, Achim Zeileis, and Kurt Hornik. Visualizing independence using extended association plots. In Kurt Hornik, Friedrich Leisch, and Achim Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*, 2003.

[101] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. YALE: Rapid prototyping for complex data mining tasks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 935–940. ACM, 2006.

[102] MA Migut, M Worring, and CJ Veenman. Visualizing multi-dimensional decision boundaries in 2d. *Data Mining and Knowledge Discovery*, pages 1–23, 2013.

[103] Malgorzata Migut and Marcel Worring. Visual exploration of classification models for risk assessment. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 11–18. IEEE, 2010.

[104] Silvia Miksch and Wolfgang Aigner. A matter of time: Applying a data-users-tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics, Special Section on Visual Analytics*, 38:286–290, 2014.

[105] Silvia Miksch and Heidrun Schumann. *Visualization of time-oriented data*. Springerverlag London Limited, 2011.

[106] Kazuo Misue. Drawing bipartite graphs as anchored maps. In *Proceedings of the Asia-Pacific Symposium on Information Visualisation (APVIS)*, pages 169–177. Australian Computer Society, Inc., 2006.

[107] Matthew Newton, Ondrej Sýkora, and Imrich Vrt'o. Two new heuristics for two-sided bipartite graph drawing. In *Graph Drawing*, pages 465–485. Springer, 2002.

[108] Chris North and Ben Shneiderman. Snap-together visualization: a user interface for coordinating visualizations via relational schemata. In *Proceedings of the working conference on Advanced visual interfaces*, pages 128–135. ACM, 2000.

[109] Tim Pattison and Matthew Phillips. View coordination architecture for information visualisation. In *Proceedings of the Asia-Pacific symposium on Information visualisation-Volume*, pages 165–169. Australian Computer Society, Inc., 2001.

[110] Tuan Pham, Ronald Metoyer, Katerina Bezrukova, and Chester Spell. Visualization of cluster structure and separation in multivariate mixed data: A case study of diversity faultlines in work teams. *Computers & Graphics*, 38:117–130, 2014.

[111] RM Pickett and GG Grinstein. Iconographic displays for visualizing multidimensional data. In *Systems, Man, and Cybernetics, 1988. Proceedings of the 1988 IEEE International Conference on*, volume 1, pages 514–519. IEEE, 1988.

[112] H. Piringer and M. Buchetics. Exploring proportions: Comparative visualization of categorical data. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 295 –296, 2011.

[113] Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. CLEF-IP 2011: Retrieval in the intellectual property domain. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[114] Margit Pohl, Florian Scholz, Simone Kriglstein, Bilal Alsallakh, and Silvia Miksch. Evaluating the dot-based contingency wheel: Results from a usability and utility study. In *Human Interface and the Management of Information. Information and Knowledge Design and Evaluation*, pages 76–86. Springer, 2014.

[115] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[116] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009.

[117] Davood Rafiei and Stephen Curial. Effectively visualizing large networks through sampling. *Visualization Conference, IEEE*, pages 375 – 382, 2005.

[118] Molham Rajjo. Evaluation of an information visualization technique for large overlapping sets. Technical Report Master's thesis, Vienna University of Technology, to be finalized in September 2014.

[119] J. N. K. Rao and A. J. Scott. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *The Journal of the American Statistical Association*, 76:221–230, 1981.

[120] Ramana Rao and Stuart K Card. The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 318–322. ACM, 1994.

[121] P. Rheingans and M. desJardins. Visualizing high-dimensional predictive model quality. In *Proceedings of IEEE Visualization*, pages 493–496, 2000.

[122] Jonathan C Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Coordinated and Multiple Views in Exploratory Visualization (CMV)*, pages 61–71. IEEE, 2007.

[123] J. O. Robinson. *The Psychology of Visual Illusion*. Dover Publications, Inc., 1998.

[124] Peter Rodgers, Leishi Zhang, and Andrew Fish. General Euler diagram generation. *Diagrammatic Representation and Inference*, pages 13–27, 2008.

[125] Jr. Rodrigues, J.F., A.J.M. Traina, and Jr. Traina, C. Frequency plot and relevance plot to enhance visual data exploration. In *Proceedings of Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2003)*, pages 117–124, October 2003.

[126] Geraldine E. Rosario, Elke A. Rundensteiner, David C. Brown, Matthew O. Ward, and Shiping Huang. Mapping nominal values to numbers for effective visualization. *Information Visualization*, 3(2):80–95, June 2004.

[127] Antonio Paulo Santos and Fatima Rodrigues. Multi-label hierarchical text classification using the ACM taxonomy. *14th Portuguese Conference on Artificial Intelligence (EPIA)*, pages 553–564, 2009.

[128] H. Schulz. Treevis.net: A tree visualization reference. *Computer Graphics and Applications, IEEE*, 31(6):11–15, 2011.

[129] Hans-Jörg Schulz, Mathias John, Andrea Unger, Heidrun Schumann, et al. Visual analysis of bipartite biological networks. In *Eurographics Workshop on Visual Computing for Biomedicine*, 2008.

[130] Hans-Jorg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. A design space of visualization tasks. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2366–2375, 2013.

[131] David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, December 1979.

[132] Michael Sedlmair, Andrada Tatu, Tamara Munzner, and Melanie Tory. A taxonomy of visual cluster separation factors. In *Computer Graphics Forum*, volume 31, pages 1335–1344. Wiley Online Library, 2012.

[133] C. Seifert and M. Granitzer. User-based active learning. In *Data Mining Workshops (ICDMW), IEEE International Conference on*, pages 418–425, 2010.

[134] C. Seifert and E. Lex. A novel visualization approach for data-mining-related classification. In *Information Visualisation (IV), 13th International Conference*, pages 490–495, 2009.

[135] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[136] Faisal Shafait, Matthias Reif, Christian Kofler, and Thomas M Breuel. Pattern recognition engineering. In *RapidMiner Community Meeting and Conference*, volume 9, 2010.

[137] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[138] Ben Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics (TOG)*, 11(1):92–99, January 1992.

[139] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[140] Harri Siirtola and Erkki Mäkinen. Constructing and reconstructing the reorderable matrix. *Information Visualization*, 4(1):32–48, 2005.

[141] P. Simonetto and D. Auber. Visualise undrawable Euler diagrams. In *12th International Conference Information Visualisation (IV)*, pages 594–599. IEEE, 2008.

[142] Paolo Simonetto, David Auber, and Daniel Archambault. Fully automatic visualisation of overlapping sets. *Computer Graphics Forum*, 28(3):967–974, 2009.

[143] Jeffrey S. Simonoff. *Analyzing Categorical Data*. Springer, 2nd edition, 2003.

[144] Markus Steinberger, Manuela Waldner, Marc Streit, Alexander Lex, and Dieter Schmalstieg. Context-preserving visual links. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2249–2258, 2011.

[145] Maureen C. Stone, Ken Fishkin, and Eric A. Bier. The movable filter as a user interface tool. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, CHI '94, pages 306–312, New York, NY, USA, 1994. ACM.

[146] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S Tan. EnsembleMatrix: Interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1283–1292. ACM, 2009.

[147] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S Tan. EnsembleMatrix: Interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, pages 1283–1292. ACM, 2009.

[148] Andrada Tatu, Georgia Albuquerque, Martin Eisemann, Jörn Schneidewind, Holger Theisel, Marcus Magnor, and Daniel Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *IEEE Symposium on Visual Analytics Science and Technology (VAST).*, pages 59–66. IEEE, 2009.

[149] David MJ Tax, Martijn Van Breukelen, Robert PW Duin, and Josef Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern recognition*, 33(9):1475–1485, 2000.

[150] Soon Tee Teoh and Kwan-Liu Ma. PaintingClass: interactive construction, visualization and exploration of decision trees. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 667–672, New York, NY, USA, 2003. ACM.

[151] James J Thomas and Kristin A Cook. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society Press, 2005.

[152] James J. Thomas and Kristin A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society, 2005.

[153] Daniel Tunkelang. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80, 2009.

[154] Cagatay Turkay, Peter Filzmoser, and Helwig Hauser. Brushing dimensions-a dual visual analysis model for high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2591–2599, 2011.

[155] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[156] Antony Unwin, Martin Theus, and Heike Hofmann. *Graphics of Large Datasets: Visualizing a Million*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[157] Tim Van de Voorde, William De Genst, and Frank Canters. Improving pixel-based vhr land-cover classifications of urban areas with post-classification techniques. *Photogrammetric Engineering and Remote Sensing*, 73(9):1017, 2007.

[158] S. van den Elzen and J.J. van Wijk. Baobabview: Interactive construction and analysis of decision trees. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 151–160, 2011.

[159] John Venn. On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(59):1–18, 1880.

[160] Anne Verroust and Marie-Luce Viaud. Ensuring the drawability of extended Euler diagrams for up to 8 sets. *Diagrammatic Representation and Inference*, pages 271–281, 2004.

[161] Tatiana Von Landesberger, Arjan Kuijper, Tobias Schreck, Jörn Kohlhammer, Jarke J van Wijk, J-D Fekete, and Dieter W Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. 30(6):1719–1749, 2011.

[162] Matthew Ward, Georges Grinstein, and Daniel Keim. *Interactive data visualization: foundations, techniques, and applications*. AK Peters, Ltd., 2010.

[163] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian Witten. Interactive machine learning: letting users build classifiers. *Intl. Journal of Human-Computer Studies*, 55(3):281–292, 2001.

[164] Chris Weaver. Building highly-coordinated visualizations in improvise. In *IEEE Symposium on Information Visualization (InfoVis)*, pages 159–166. IEEE, 2004.

[165] M. Wertheimer. Laws of organization in perceptual forms. In W. D. Ellis, editor, *A sourcebook of Gestalt psychology*, pages 71–88. Routledge and Kegan Paul, 1938.

[166] L. Wilkinson. Exact and approximate area-proportional circular Venn and Euler diagrams. *Visualization and Computer Graphics, IEEE Transactions on*, 18(2):321–331, 2012.

[167] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[168] K. Wittenburg, A. Malizia, L. Lupo, and G. Pekhteryev. Visualizing set-valued attributes in parallel with equal-height histograms. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI)*, pages 632–635. ACM, 2012.

[169] Kent Wittenburg, Tom Lanning, Michael Heinrichs, and Michael Stanton. Parallel bargrams for consumer-based information exploration and choice. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 51–60. ACM, 2001.

[170] David F. Wyatt. `http://www-edc.eng.cam.ac.uk/tools/set_visualiser/`. accessed: August 2014.

[171] Tengke Xiong, Shengrui Wang, Andre Mayers, and Ernest Monga. A new MCA-based divisive hierarchical algorithm for clustering categorical data. In *Proceedings of IEEE International Conference on Data Mining*, pages 1058–1063. IEEE Computer Society, 2009.

[172] Lei Xu, Adam Krzyzak, and Ching Y Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(3):418–435, 1992.

[173] Ji Soo Yi, Rachel Melton, John Stasko, and Julie A Jacko. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256, 2005.

[174] Jianting Zhang and L. Gruenwald. Opening the black box of feature extraction: Incorporating visualization into high-dimensional data mining processes. In *IEEE International Conference on Data Mining (ICDM)*, pages 1188–1192, 2006.

[175] Jian Zhao, Christopher Collins, Fanny Chevalier, and Ravin Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2080–2089, 2013.

[176] Lanbo Zheng, Le Song, and Peter Eades. Crossing minimization problems of drawing bipartite graphs in two clusters. In *Proceedings of the Asia-Pacific Symposium on Information Visualisation (APVIS)*, pages 33–37. Australian Computer Society, 2005.

Curriculum Vitae

Homepage	http://www.cvast.tuwien.ac.at/~bilal

Work experience

01.06.2010 – Now	Research Assistant at Vienna University of Technology, Austria
01.09.2009 – 31.01.2012	Software Developer at AMOS-Austria
05.08.2008 – 30.06.2009	Research assistant at UC4 Software GmbH, Vienna, Austria
13.08.2007 – 30.09.2007	Internship at VRVis Forschungs GmbH, Vienna, Austria

Education

October 2010 – Sept 2014	PhD candidate in Computer Science at Vienna University of Technology
Sept 2010	The UKVAC Visual Analytics Summer School, Middlesex University London.
March 2007 - July 2009	Master of Science (MSc.) at Vienna University of Technology
Sept 2001 - July 2006	Bachelor of Science (BSc.) at Damascus University – Faculty of Information Technology

Research Interests

Visual Analytics and Information Visualization
Set Visualization and Search interfaces
Pattern Recognition and Machine Learning
Software Visualization

Award and Honors

2012	Best Paper Honorable mention at IEEE Conference on Visual Analytics Science and Technology
2008	TU Vienna PRIP Prize (Pattern Recognition and Image Processing)
2006	Damascus University Best Bachelor Graduate Award

Services to the Scientific Community

2014	Organizing Committee of the IEEE VIS Conference.
2013-2014	Program Committee of the IEEE Working Conference on Software Visualization - "NIER" and "Tool Demos" tracks
2010-2014	Various reviews to major conferences and journals on visualization

Invited Talks

2013	State-of-the-art methods for visualizing set-typed data *Invited Seminar Talk at the University of Kent - School of Computing*