

Dipl.-Ing. Belgin Mutlu, BSc

# Methods to Recommend and Personalize Visualizations

---

Dissertation  
submitted to  
Graz University of Technology



for the attainment of the degree of  
Doctor of Engineering Sciences (Dr.techn.)

Institute of Interactive Systems and Data Science  
Graz University of Technology

First supervisor: Univ.-Prof. Dipl.-Inf. Dr. techn. Stefanie Lindstaedt  
Second supervisor: Ass. Prof. Denis Parra Santander  
Advisor: Dr. techn. Eduardo Veas

Graz, June 2018

## EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen / Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am .....  
(Unterschrift)

## STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, the .....  
(Signature)

## Kurzfassung

Die Datenanalyse (en.: data analysis) und speziell das Data-Mining (en.: data mining) haben als aufkommende Disziplinen in den letzten Jahren zunehmend an Bedeutung sowohl in der Akademie als auch in der Industrie gewonnen. Die Forschungsinitiativen wie Industry 4.0, SmartFactory oder Internet-of-Things (IoT) sind Beispiele dafür, die Forscher und die Ingenieure dazu zu bringen, die Produktion und Produkten in diversen Anwendungsgebieten zu verbessern, indem sie verschiedene Technologien aus dem Gebiet der Datenanalyse ausschöpfen und optimal einsetzen. Neben vielen Herausforderungen, die dieses Vorhaben mit sich bringt, stellt das ständige Wachstum der Daten/Informationen (das sogenannte Information-Overload Problem (en.: information overload problem) das Hauptproblem dar. Einer der effektiven Wege dieses Overload Problem zu bewältigen ist die visuelle Datenanalyse.

Visualisierungen weisen spezielle Vorteile bei der Behandlung vom Information-Overload Problem vor, da sie als integrale Aspekte der visuellen Wahrnehmung der Menschen betrachtet werden können. Mit anderen Worten, viele Menschen erkennen sofort, was Ihnen visuell angeboten wird. Die weiteren Erkenntnisse wiederum können einfach durch die visuellen Operationen gewonnen werden, die Menschen auf Grund ihrer visuellen Gegebenheiten auf natürliche Weise durchführen können, sowie etwa das Clustering, die Filterung und vergleichen der verschiedenen Größen (en.: comparing quantities). Auf der anderen Seite ist die Erstellung entsprechender Visualisierungen ein komplexes Unterfangen, welches ein adäquates Wissen aus dem Anwendungsgebiet sowie Wissen über die Daten erfordert. Dies bedeutet weiters, dass ein gewöhnlicher Benutzer, der keine Expertenkenntnisse vorweisen kann (en.: non-expert user), sich schwer dabei tun wird, die richtigen Visualisierungen für seine Daten zu erzeugen. Dieses Problem leitet somit das erste Forschungsziel dieser Dissertation ab: basierend auf sogenannten visuellen Kodierungsregelungen und Wahrnehmung -Richtlinien (en.: visual encoding rules and perception guidelines) wurde eine Methode entwickelt, mit der die adäquaten Visualisierungen automatisch aus der Datenmenge generiert werden können. Als Endergebnis wird die visuelle Datenanalyse auch für gewöhnliche Benutzer möglich gemacht.

Die automatisierte Generierung der Visualisierungen, bei dem nur die visuellen Kodierungsregelungen befolgt werden, sowie es die Fachliteratur vorschlägt [MHT<sup>+</sup>14], bringt große Einschränkungen mit sich. Das Hauptproblem hierbei ist, dass diese Vorgehensweise zu einer Vielzahl von Visualisierungen führt. Dabei sind die erzeugten Visualisierungen hinsichtlich der Darstellung der Daten korrekt, jedoch zeigen

sie nicht immer was der Benutzer sehen möchte, sei es der Visualisierungstyp an sich oder die Informationen, die die Visualisierung abgebildet. Daher wurde im Rahmen dieser Dissertation ein innovatives Recommender System entworfen, welches (i) die Visualisierungen unter der Berücksichtigung der visuellen Kodierungsregelungen (en.: perception rules) erzeugt und (ii) diejenigen Visualisierungen herausfiltert, die mit den Präferenzen des Benutzers (en.: user preferences) bestens übereinstimmen. Dabei wurden verschiedene Strategien (Verfahren) eingesetzt und evaluiert. Zum einen wurde eine Multi-Dimensionale Skala von Ratings definiert, um das sogenannte Collaborative Filtering Verfahren zu evaluieren. Zum anderen wurden Tags verwendet, um ein Recommendation Verfahren einzuleiten, welches unter den Namen Content-based filtering bekannt ist. Die Tags repräsentieren dabei die Interessen des Benutzers bezüglich der Informationen, die er visuell betrachten möchte.

Letztendlich wurde ein hybrides Recommender System (en.: hybrid filtering) entworfen, das diese beiden Arten von Informationen kombiniert, d.h. was die Visualisierung (Tags) zeigt und wie gut sie ist (Ratings), um visuelle Empfehlungen (en.: visual recommendations) zu generieren. Mit diesem letzten Schritt wurde ein vollständiges Framework für personalisierte visuelle Datenanalyse entworfen.

Warum ein Benutzer eine bestimmte Visualisierung bevorzugt hängt von vielen Faktoren ab. Es kann sein, dass der Benutzer damit eine bestimmte Aufgabe oder ein bestimmtes Problem lösen möchte. Oder es sind die visuellen Operationen, die die jeweiligen Visualisierungen anbieten und die Information, die der User aus den Visualisierungen extrahieren kann, die dazu führen, dass eine bestimmte Visualisierung von einem Benutzer mehr bevorzugt wird als eine andere. Wichtig dabei ist, wie der Benutzer seine Präferenzen ausdrückt. Das dritte Ziel dieser Dissertation daher war (i) die Information Quellen, die der Benutzer benutzt um seine Präferenzen zu teilen, auf ihre Eigenschaften zu untersuchen und (ii) die bestmögliche Methode zu finden, welche diese Informationen benutzt um visuelle Empfehlungen (en.: visual recommendations) zu generieren.

Zu diesem Zweck wurden folgende vier Quellen untersucht: Ratings, Tags, Titel (en: titles) und Fragen&Antworten (en: questions&answers). Im ersten Schritt wurde die Qualität von Ratings mittels statistischen Methoden analysiert. Im weiteren Schritt wurde unter Verwendung von Messverfahren aus der Informationstheorie die Qualität von textuellen Quellen (Tags, Titel und Fragen&Antworten) evaluiert. Die Anwendbarkeit aller vier Quellen wurde anschließend mit den integrierten Recommendation Verfahren untersucht.

Die Arbeit in dieser Dissertation wurde mit acht Publikationen begleitet und trägt wissenschaftlich zum Forschungsgebiet "Visuelle Empfehlungsdienste" (en: visualization recommendation) bei. Die beschriebene Methode in dieser Dissertation stellt einen neuen Ansatz vor, mit dem der gesamte Visualisierungsprozess automatisiert und personalisiert werden kann. Dieser Prozess wird dadurch optimiert, dass die wesentlichen Interessen bzw. Präferenzen von Benutzern berücksichtigt werden.

Die einzelnen Beiträge dieser Dissertation können wie folgend zusammengefasst werden:

Erstens wurde eine Visualisierungs-Ontologie (en.: visualization ontology), die das gesamte Experten-Wissen über Visualisierungen aufnehmen kann, definiert. Dabei wurde eine ausführliche Analyse vom Stand der Technik durchgeführt. Die definierte Ontologie basiert auf einigen bekannten Arbeiten im Gebiet der Visualisierung: die Prinzipien und bewährte Methoden von Tamara Munzer, Bertins Theorie von Visualisierungen der Informationen und die Generierung graphischer Präsentationen von Mackinlay.

Zweitens wurde ein regeln-basiertes Recommender-System entworfen, das diese Ontologie verwendet, um Visualisierungen automatisch zu generieren. Im Vergleich zu den existierenden Ansätzen im Gebiet, arbeitet dieses Recommender-System vollständig autonom. Was so viel heißt wie, dieses System verlangt von den Benutzern keine manuelle Spezifikation, die definiert, wie die Daten auf Visualisierungen dargestellt werden müssen.

Drittens wurde ein personalisiertes Recommender-System entworfen, das personalisierte visuelle Empfehlungen definiert, in dem er ein Katalog von Benutzer-Präferenzen verwendet. Um verschiedene Typen von Benutzer-Präferenzen im Rücksicht zu nehmen (Ratings, Tags), verwendet dieses Recommender-System verschiedene Methoden für die Empfehlungen: collaborative filtering, content-based filtering und ein hybrid filtering.

Viertens wurde ein Evaluierungs- Framework definiert, mit dem es ermöglicht wurde, die Qualität von verschiedenen Typen von Benutzer Präferenzen zu evaluieren und somit die richtigen Schlussfolgerungen zu ziehen.

## Abstract

Data analysis and data mining in particular as emerging fields of data science have gained wide popularity in the last years both in academia and in industry. As an example, recent research initiatives such as Industry 4.0, SmartFactory or Internet-of-Things (IoT) try to motivate researchers and engineers to improve the production and products in various application fields by utilizing technologies for data analysis such as clustering, filtering and visual data analysis for example. Beside many issues that have to be solved in this intent, the major problem that still remains here is how to deal with the growing amount of data/information, i.e., a so-called information overload problem. Visual data analysis has been proven to be one of the effective ways to tackle this problem.

Visualizations have a distinctive advantage when dealing with the information overload problem: because they are grounded in basic visual cognition, many people understand them and can naturally perform visual operations such as clustering, filtering and comparing quantities. However, creating appropriate visualizations requires specific expertise of the domain and underlying data. Yet, an ordinary user lacks expert knowledge and can rarely generate sophisticated visualizations. Thus, the first quest of this dissertation was to provide strategies to automatically recommend appropriate visualizations for non-experts by following visual encoding rules and perceptual guidelines.

Yet, considering just visual encoding rules proposed in the literature [MHT<sup>+</sup>14] leads to a large set of possibilities, valid in terms of representing the data visually, but without considering which type serves the users' needs best. To tackle this issue, we proposed a novel recommender system that (i) recommends visualizations based on a set of visual cognition rules and (ii) filters a subset considering the user's preferences.

This dissertation investigated different strategies to recommend visualizations considering different aspects of the user preferences. A multi-dimensional scale is used to estimate aspects of quality for visualizations for collaborative filtering. Alternatively, tag vectors describing visualizations are used to recommend potentially interesting visualizations based on content. Finally, a hybrid approach combined information on what a visualization is about (tags) and how good it is (ratings).

A user places interests on several aspects of a visualization, the task or problem it helps to solve, the operations it permits, or the features of the dataset it represents. This dissertation further concentrated on characterizing the user preferences, in particular: i) the sources of information used to describe the user's preferences/needs

and the content of visualizations and ii) the methods to produce the most suitable recommendations thereby.

We considered four information sources corresponding to different aspects of interest: ratings, tags, titles and questions&answers. First, we investigated the quality of ratings in representing the user preferences applying statistical analysis tools. Next, using information-theoretic measures we investigated the efficiency of text-based inputs (tags, titles, questions&answers) in describing the user preferences and the visualization content, the user and item models respectively. Finally, the practicability of each input is evaluated by applying the three recommendation strategies.

This dissertation resulted in eight publications and contributes scientifically in the research area of visualization recommendation. It provides a novel approach on how to automate and personalize the visualization process in order to better address the characteristics of the user's data and, more important, the user's visual preferences, needs and interest.

The contributions of this dissertation can be summarized as follows:

A visualization ontology has been defined to collect expert knowledge about visualizations in a common persistence format. This required an extensive background and state of the art analysis. We mainly focused on Tamara Munzner's work which captures principles and best practices of information visualization, Bertin's theory on information visualization and Mackinlay's formal language to generate graphical presentations.

Second, a rule-based visual recommender has been developed which uses the visualization ontology to automatically define and create appropriate visualizations for a given dataset. In contrast to the existing work, our visual recommender is completely autonomous: there is no need for manual specification of visual information which basically requires specific expertise in visualization principles.

Third, a personalized visual recommender has been defined which uses a catalog of the user preferences to define personalized visual recommendations. To achieve that, the recommender applies different methods, collaborative filtering, content-based filtering and hybrid filtering respectively, each of which applies a different kind of the user input (ratings, tags) describing the user's visual needs and interest.

Fourth, we proposed a framework to assess the encoding power of different information sources in describing the user preferences and the visualizations.

## Acknowledgements

Writing this thesis was only possible with the support and collaboration of many great people. First of all, I would like to say thank you to my supervisor Stefanie Lindstaedt for the advices and guidance throughout the different stages of this thesis. My greatest thanks go to Eduardo Veas for his invaluable support and inspiring guidance. Without his excellent advices and continuous support, this work would not have been possible. I am grateful that he accepted the role of being my mentor.

My special thanks go to Vedran Sabol who supported me with his extensive domain knowledge in the past years. I want to further thank Patrick Höfler, Gerwald Tschinkel, Christoph Trattner and Ilija Simic for the excellent collaboration and support throughout the different stages of this thesis. Further, I would like to thank my colleague and friend Cecilia di Sciascio who helped me to keep motivated and focused during the last years of this dissertation. In addition, I would like to thank the members of the Knowledge Visualization group for the inspiring working atmosphere and their support.

Finally, I would like to thank Nermin, my mother, my father and my brother for supporting and motivating me in the past years.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Research Context . . . . .	2
1.2	Research Questions . . . . .	3
1.3	Scientific Contribution . . . . .	4
1.4	Research Papers . . . . .	6
1.5	Structure of the Thesis . . . . .	10
1.6	Summary . . . . .	11
<b>2</b>	<b>Related Work</b>	<b>13</b>
2.1	Background: Recommending Visualizations . . . . .	13
2.1.1	Information Visualization . . . . .	14
2.1.2	Semantic Web . . . . .	20
2.1.3	Recommender Systems . . . . .	25
2.1.4	Summary . . . . .	29
2.2	State-of-the-art: Recommending Visualizations . . . . .	30
2.2.1	Modeling Visual Representations . . . . .	31
2.2.2	Visualization Recommendation . . . . .	38
2.2.3	Crowd Sourcing Visualization Studies . . . . .	50
2.3	Summary . . . . .	51
<b>3</b>	<b>Automated Visualization Recommendation</b>	<b>53</b>
3.1	Visual Analytics Vocabulary . . . . .	53
3.1.1	Formalization of Visualization Knowledge . . . . .	53
3.1.2	Visualization Ontology . . . . .	56
3.1.3	Visual Pattern . . . . .	56
3.1.4	Summary . . . . .	58
3.2	Rule-based Recommender . . . . .	58
3.2.1	System Overview of the Rule-based Recommender . . . . .	59
3.2.2	Usage Scenario: Automated Visualization of Scientific Data . . . . .	69
3.2.3	Evaluation . . . . .	70
3.2.4	Summary . . . . .	75

<b>4</b>	<b>A Crowd-sourced Study for Preference Elicitation</b>	<b>77</b>
4.1	Experiment Setup . . . . .	77
4.1.1	Datasets . . . . .	77
4.1.2	Procedure . . . . .	78
4.2	Results . . . . .	80
4.2.1	Participants . . . . .	80
4.2.2	Variability in Preference Scores . . . . .	81
4.3	Summary . . . . .	84
<b>5</b>	<b>Personalized Visualization Recommendation</b>	<b>87</b>
5.1	Approach . . . . .	87
5.1.1	Collaborative Filtering . . . . .	89
5.1.2	Content-based Filtering . . . . .	91
5.1.3	Hybrid Filtering . . . . .	94
5.2	Use-Case: Visual Recommender for Learning in Communities . . . . .	95
5.3	Evaluation . . . . .	98
5.3.1	Evaluation Protocol . . . . .	98
5.3.2	Recommendation Quality . . . . .	100
5.3.3	Discussion . . . . .	107
5.4	Summary . . . . .	108
<b>6</b>	<b>Content Descriptors for Visual Recommender Systems</b>	<b>111</b>
6.1	Related Work . . . . .	113
6.2	Experiment Setup . . . . .	114
6.3	Encoding power of User-provided Input . . . . .	115
6.3.1	Methodology . . . . .	115
6.3.2	Results . . . . .	116
6.4	Recommendation Quality . . . . .	120
6.4.1	Content Based Recommender . . . . .	121
6.4.2	Hybrid Recommender . . . . .	123
6.4.3	Discussion . . . . .	124
6.5	Summary . . . . .	125
<b>7</b>	<b>Conclusion</b>	<b>127</b>
7.1	Research Question . . . . .	127
7.2	Self- Reflection about this Thesis: Lesson Learned . . . . .	130
7.3	Future Work . . . . .	134
7.4	Visualization Ontology & Rule-based Visual Recommender . . . . .	135
7.5	Personalized Visual Recommender . . . . .	135
7.6	Summary . . . . .	136

<b>8 Appendix</b>	<b>137</b>
8.1 Evaluating Rule-based Recommender System . . . . .	137
8.2 Amazon Mechanical Turk Study . . . . .	142
8.2.1 Demographics . . . . .	144
8.3 Visual Analytics Vocabulary . . . . .	146
8.4 Visualization Ontology . . . . .	147
<b>References</b>	<b>160</b>



# List of Figures

2.1	In a flat table an item is represented by a row and an attribute by a column. Their intersections represent a cell which contains the value of the pairwise combination of a row and a column (from [Mun14], Figure 2.5). . . . .	15
2.2	Table with the attribute columns. The columns are color coded regarding to their type: quantitative, ordinal and categorical (from [Mun14], Figure 2.9). . . . .	16
2.3	There exist four types of marks: point, line, area and volume (from [Mun14], Figure 5.2). . . . .	17
2.4	Visual channels. There are four position channels: aligned planar position, unaligned planar position, depth and spatial region. Other visual channels are represented by color with its three aspects: hue, saturation and luminance. There exist three size channels: length in one-dimensional size, area in two-dimensional size and volume in three-dimensional size. Further visual channels are tilt (also called angel) and shape (from [Mun14], Figure 5.3). . . . .	18
2.5	Channel ranking according to their effectiveness in encoding data attributes. Note, magnitude channels should be used for ordered data and identity channels for the categorical data (from [Mun14], Figure 5.6). . . . .	19
2.6	Sample triples: a list of plays that Shakespeare wrote or a list of parts of the United Kingdom. From [AH11], Table 3.3. . . . .	21
2.7	High level architecture of Content-based recommender. From [RRSK10], Figure 3.1. . . . .	28
2.8	XML File presenting a visualization including the data it shows. From [Wol15], Figure 3. . . . .	32
2.9	GraphXML describing basic components of a graph, its nodes and edges respectively. From [KKMN10], Figure 5. . . . .	33
2.10	A fragment of the Visualization Ontology describing the data model. From [SAR08], Figure 2. . . . .	34
2.11	The Visual Representation Ontology illustrating a 2D Graph for music charts. From [GSGC08b], Figure 4. . . . .	35
2.12	Diagrammatic view of SGO (see left side) describing a line graph showing the yearly percentage change in e.g., manufacturing sales (see right side). From [DFVR10], Figure 3 & 4. . . . .	36

2.13	The Polaris user interface. To create a visualization, the user first selects the desired fields from the database schema onto shelves. The system then performs the required visual operations to visually display the values of the selected fields. From [STH02], Figure 1. . . . .	40
2.14	A map showing the residences of the U.S senators. From [CDC <sup>+</sup> 07], Figure 2. . . . .	41
2.15	The ShowMe user interface. For creating a visualization, the user first drags the fields of interest from so called data window on the so called shelves (row, column), which organizes the fields regarding to their type: in dimensions or measures respectively. After this, the system selects the most appropriate visualization for the selected fields and presents the data using best practices of graphic design. From [MHS07], Figure 1. . . . .	42
2.16	The Vispedia Workflow. On a certain Wikipedia page the user first selects a table using the Vispedia bookmarklet. As next, she picks one of the three visualizations for which she has to formulate an initial search query that comprises the appropriate fields for the selected visualization. Once the system detects the data matching the query, it finally creates the candidate visualization. From [CWT <sup>+</sup> 08], Figure 2. . . . .	43
2.17	Many Eyes Schema for visual mapping. From [VWV <sup>+</sup> 07], Table 1. . . . .	44
2.18	SEEDB: (a) Shows user’s current visualization. (b) Shows the visualization that has been recommended by the system since it shows data that deviate from them shown in (a). (c) Shows a visualization that would not be recommended since it follows a similar trend like (a). From [VMPP14], Figure 1, 2 and 3. . . . .	45
2.19	The Voyager user interface: The schema panel (left) lists the user-selected attributes. The main gallery (right) presents recommended visualizations of different attribute subsets and transformations. From [WMA <sup>+</sup> 16], Figure 1. . . . .	46
2.20	The adaptive VIBE showing the document placements based on their similarity to the queried term and the user model of the current user. The yellow circles illustrate the queried term whereby the blue circles denotes the extracted named-entities and represent the user model. The squares represent the resulted documents whereby their position indicate their similarity to the queried term or user model. From [AB09], Figure 2. . . . .	47
2.21	Behavior-driven visualization recommender. To use the recommender, the user first defines a query (a). Next, the system generates a visualization the user can interact with to analyze her data (b). Based on user’s current behavior, the system defines visual recommendations, notify (c) the user via “magic wand” icon in the history panel and (d) a flashing segment on the recommendation sidebar. The user can either accept the recommendation by clicking on the icon or deny it. From [GW09], Figure 1. . . . .	48
3.1	Incorrect Bar chart encoding two categorical data. From [CMS99], Figure 1.25. . . . .	54
3.2	Visual patterns for the bar chart and timeline defined in the description vocabulary. . . . .	58

3.3	Recommendation Dashboard using VizRec to automatically recommend appropriate visualizations. On the left side, the list with the recommendations is shown. On the right side one of the recommend visualization is shown which the user uses to filter the recommendations regarding to their language (English, German, unknown). . . . .	59
3.4	Schematic representation of the rule-based recommender: The stages (a), (b) and (c) illustrate the preprocessing unit. The stage (d) illustrates the visual mapping process between the elements and the visual patterns, whereby the defined mapping combinations are shown in stage (e). Finally, the recommendations will be presented to the user in a random order (f). . . . .	60
3.5	Preprocessing: The input data (a) for VizRec are structured following a metadata model (b). The preprocessing unit is responsible, first, for the identification and extraction of the metadata elements and second, for the datatype analysis on the values of the metadata elements (c). . . . .	62
3.6	Visual mapping process. . . . .	64
3.7	Some of the bar chart combinations generated for the dataset <i>movies</i> using the bar chart patterns (1), (2), (3) and (4). The patterns are listed in Table 3.1. . . . .	65
3.8	Schematic representation of the CODE pipeline to extract and analyze scientific facts embedded in publications. . . . .	69
3.9	Pipeline to extract and structure scientific facts embedded in publications. . . . .	70
3.10	The Query Wizard serves as an interface to the Linked Data endpoints. Once a RDF Data Cube is queried from the endpoint, Query Wizard presents the graph structure of the Cube content in a tabular form. Using the operations (aggregation, filtering, removing/adding columns) provided with the table, user can perform various exploration tasks on her data before passing it to VizRec. . . . .	71
3.11	The user interface of VizRec's rule-based recommender. Once Query Wizard passes the RDF Data Cube to VizRec's rule-based recommender, the recommender extracts the dimensions and measures (see bottom left) and defines the list of appropriate visualizations (top left). Once the user selects one of the selected visualizations, the visualization will be automatically generated (right). . . . .	72
3.12	Subjective Workload measured using NASA Task Load Index. . . . .	74
4.1	Crowd-sourced Experiment Task (HIT). Participants were motivated to carefully observe the visualization with the study task, in terms of writing tags and a title. Thereafter, they had to rate it in a multidimensional scale and pose a question that is answered with the visualization. . . . .	79
4.2	Three (in average) highest rated visualizations for the datasets <i>Movies</i> , <i>EU</i> and <i>Books</i> including the five most frequently used tags. . . . .	83
4.3	Three (in average) lowest rated visualizations for the datasets <i>Movies</i> , <i>EU</i> and <i>Books</i> including the five most frequently used tags. . . . .	84

4.4	Mean and variability in scores (rating 1–7, higher is better). The heatmap illustrates the contribution of 9 dimensions (US=useful, SA=satisfying, EF=efficient, UN=Easy to understand, co=confusing or=organized, cl=cluttered, EX=exciting, bo=boring) to the overall score (SC). The violin plot below illustrates the high variability in personal ratings. . . . .	85
4.5	Coefficient of variation. Overall variation for all charts (top), variation broken down by chart type (middle) and by dataset (bottom). Note, that coefficient of variation is the ratio of SD/Mean. The density in the violin chart shows where the broad variation of scores across charts. . . . .	85
4.6	Significant Interactions Visualization Type / dataset. The heat-map illustrates the mean score and standard deviation for each combination of <i>dataset-visualization type</i> (1=completely disagree, 7=totally agree). The lines below show where differences begin to be significant. Note, that due to its high variability, <i>books-bar</i> is not significantly better than <i>eu-line</i> , whereas <i>movies-line</i> is. . . . .	86
5.1	Schematic representation of the VizRec recommendation pipeline: The stages (a), (b) and (c) illustrate the preprocessing unit. The stage (d) illustrates the visual mapping process between the elements and the visual patterns, whereby the defined mapping combinations are shown in stage (e). For the personalized visualization recommendation VizRec uses either the user ratings, user and item profiles or a combined version of both (f). Finally, the recommendations will be presented to the user in a top-n manner (g). . . . .	88
5.2	Personalized learning dashboard: Teacher explores the number of previous attempts of females and males who take her class with regard to their education level. The teacher first selects the data attributes of interest (see left) the system then recommends the personalized visualizations (see right). One selected, the personalized visualizations are created automatically (see center). . . . .	98
5.3	Performance plots for collaborative filtering (CF), content based (CB) and hybrid approaches for different sizes of training sets (20%-80%) considering the first three recommendations in the result list ( $k = 3$ ). . . . .	109
6.1	Schematic representation of the visualization recommender: The rule-based stage applies visual encoding guidelines to generate a collection of visualizations appropriate for the data. The personalization stage applies the user preferences/profiles (ratings, content terms such as tags and titles) and filters the visualizations according to users' needs and interests. This stage also maintains repositories for the user preferences/profiles. . . . .	112
8.1	Amazon Mechanical Turk Evaluation: instructions. Instructions clearly stated each step of the study. The instructions page also included demographical data. . . . .	142



8.2	Amazon Mechanical Turk Evaluation: tutorial. The tutorial showed an exemplary task, with sample responses for tags, title, ratings and question&answer. . . . .	143
8.3	Amazon Mechanical Turk Evaluation: first visualization to evaluate. . . .	144
8.4	Amazon Mechanical Turk Evaluation: second visualization to evaluate. . .	145
8.5	Amazon Mechanical Turk Evaluation: third visualization to evaluate. . . .	146



# List of Tables

2.1	Life expectancy broken down by region (unitary authority), age and time. From [CRT13], Table 1. . . . .	23
3.1	Mapping combinations defined for the exemplary dataset <i>movies</i> using bar chart’s visual patterns (1)-(4). The visualizations shown in the last columns are generated for the first mapping combination of each pattern, to give an example for instantiated mapping combinations. . . . .	68
4.1	Basic statistics of the three rating and tag datasets collected via the crowd-sourced experiment. . . . .	81
4.2	Top-10 tags in each of the three datasets. The tags were stemmed (normalized) as described in Section 5.1. . . . .	82
5.1	Input data for the calculation of the k-nearest neighbors and generating predictions for the active user, including item-id, user-id and ratings. . . . .	89
5.2	An excerpt of the <i>movies</i> dataset with generated user and item profiles. . . . .	94
5.3	Recommendations for the weighted hybrid recommender. . . . .	95
5.4	Quality metrics values estimated for the three example datasets using VizRec’s CF-RS. The values are calculated first for the ratings taken from one out of the nine dimensions (bo=boring, cl=cluttered, co=confusing, EF=efficient, EX=exciting OR=organized, SA=satisfying, UN=Easy to understand, US=useful) and then for the overall rating score (O.a). Note, we inverted the ratings of the negative dimensions (N) boring, cluttered and confusing using their opposites (P) exciting organized and easy to understand according to the equation $\frac{(P-N)+7}{2}$ . For this test we used a 5-fold cross validation whereby each iteration used 80% of user’s data as training set and 20% as test set. . . . .	102
5.5	Quality metrics values R@3, P@3, F@3, MAP@3, nDCG@3 estimated for the three datasets using the baseline algorithms MP and RD ( $k = 3$ ). Note, for this test we executed a 5-fold cross validation whereby each iteration used 80% of user’s data as training set and 20% as test set. . . . .	103
5.6	Comparing VizRec hybrid approach with rating-based and tag-based approaches: Quality metric values considering the first three recommendations in the list ( $k = 3$ ). The results are distributed over four sets (from 20% to 80%), each containing different number of items (visualizations) in the training set <i>ts</i> . Note, for this test we used a 5-fold cross validation. . . . .	105

6.1	Basic statistical properties of the datasets collected via Amazon Mechanical Turk. Column "User/Vis" shows the average number of the user assigned to a visualization; "Vis/User" is the average number of visualizations assigned to a user. Note that the values in parenthesis indicate number of unique terms. . . . .	116
6.2	Example distributions of top-5 terms for the EU dataset. Note, the terms are stemmed using Porter stemmer [KG13]. . . . .	117
6.3	Information-theoretic measures for tags, titles and questions used for the user profiles. Note that the measures have been calculated among all three datasets. . . . .	118
6.4	Information-theoretic measures for tags, titles and answers used for item profiles. Note that the measures have been calculated among all three datasets. . . . .	119
6.5	The performance of our individual content based filtering approaches (CB), compared with baseline algorithm RD: quality metric values considering the first three recommendations in the list (k=3). ***Significant at $p < 0.001$ . .	122
6.6	The performance of our individual hybrid filtering approaches (HB), compared with baseline algorithm MP: quality metric values considering the first three recommendations in the list (k=3). ***Significant at $p < 0.001$ . .	124



# Chapter 1

## Introduction

In today's digital age we observe an increasing number of websites, social- and learning platforms and different computer-mediated communication systems. The massive amount of data being generated in such environments can be quickly overwhelming for the analyst, regardless from which discipline she is. As a result, relevant, useful information may be overlooked, possible links between information not identified and wrong conclusions drawn.

Visualizations have shown to be effective in dealing with huge datasets: since they are grounded on visual cognition, people understand them and can naturally perform visual operations such as clustering, filtering and comparing quantities. Concretely, by assigning data to visual characteristics people can intuitively extract valuable information and perceive properties of their data which might remain undiscovered by other means of analyzing them.

Creating appropriate visualizations, however, is challenging and time consuming, as it involves many steps. It is a complex process which requires domain knowledge, understanding of the data and knowledge about the task and the user preferences. Yet, an ordinary user is not an expert in visualizations and can rarely generate sophisticated visualizations. Still, popular visualization tools require manual specification of visual information, which involves the following: selecting variables of interest, selecting transformations and designing visual encoding.

To date, only a few approaches have attempted to automatically generate and recommend visualizations starting just from data. These approaches follow so called visual encoding rules and perceptual guidelines to define which visualization is appropriate for the given data. In this context, these rules basically are formal specifications for what might be achieved when representing data visually. While these approaches are successful in reducing the knowledge barrier, they fall short in being domain independent and, more important, they ignore the fact that the choice of visualization is a personal matter.

This thesis was motivated by the questions on how to automate the visualization process for the non expert users across various domains and how to personalize it in order to define visualizations that better address user's preferences, needs and inter-

ests. To that end, a two-stage recommender system, called VizRec, is introduced. In the first stage, rule-based stage, VizRec defines a set of appropriate visualizations using so called Visual Analytics Vocabulary that represents visualizations in a common persistence model. The produced visualizations are perceptually valid, but too general. In order to investigate whether different user prefers different visualizations, we performed a crowd-sourced study and collected user preferences and descriptions in a systematic manner. Basically, we used the collected data (i) to understand the variability in choice of preferred visualizations and (ii) to analyze the perceived quality. Yet, the results revealed that the choice of visualization is a personal matter and only a personalized visual recommender can account for such variability. Thus, in the second stage of VizRec, personalization stage, the system filters a subset considering the user’s preferences needs and interest by applying individual recommendation strategies. The goal of the this stage is to generate only the visualizations the current user would select as part of her analysis workflow. Therefore, we investigated which information about the user is needed to anticipate her choice of visualizations for data analysis and how to use this information for recommendation purposes.

First a collaborative filtering approach is introduced which uses on a multidimensional rating scale to gauge different quality aspects of the visualizations recommended. Next, the effects these aspects have on the filtering of visualizations are analyzed. Content-based filtering has been investigated as a means to recommend visualizations according to the information needs of the user by collecting the user provided tags. A combination of both aspects, collaborative- and content-based filtering, is studied using a hybrid recommendation strategy.

In the context of visualizations, the user provided input can take other forms, beside ratings and tags. For instance, it is common for a user to pose a question that is answered with a visualization, or to define a title and description for the visualization in form of a caption. We considered these two alternative sources of information (titles, questions&answers) as potential descriptors of both user intention and item features. To investigate the characteristics of these information sources in describing the user and the visualizations, information theoretic measures have been applied. This investigation should reveal how capable each of these input types is in encoding the user and the visualizations and thus are appropriate candidates to be used for visual recommender systems.

## 1.1 Research Context

The research presented in this thesis was conducted within the EU Projects *CODE-Commercially Empowered Linked Open Data Ecosystems in Research*, *EEXCESS-Enhancing Europe’s Exchange in Cultural Educational and Scientific Resources* and *AFEL- Analytics for Everyday Learning* in close cooperation with Patrick Höfler, Vedran Sabol, Gerwald Tschinkel, Ilija Simic and Eduardo Veas from the Know

Center and Christoph Trattner from University of Bergen. With Gerwald Tschinkel we worked closely together on developing the user interface for the rule-based recommender described in Section 3 and on preparing and executing the user evaluation described in Section 3.2.3. Patrick Höfler provided the infrastructure blocks (see Section 3.2.2) needed to access the data provided by project partners. Ilija Simic was the developer of the user interface for personalized visual recommender. Vedran Sabol had an important role by providing domain specific knowledge and by assisting in definition of usage scenarios. Christoph Trattner provided insightful guidance on the design of the evaluations described in Section 5.3 and in Section 6.2. Eduardo Veas acted as a supervisor for all mentioned activities and had an invaluable contribution in designing, performing and analyzing the evaluations described in Section 5.3 and in Section 6.2. My focus in this thesis was to define the research direction (with the support of Eduardo Veas) and to develop, implement and evaluate the visual recommender, called VizRec.

This thesis has been performed as part of the work for the Know-Center GmbH Research Center for Data-Driven Business&Big Data Analytic.

## 1.2 Research Questions

This thesis is motivated by the question, how to provide meaningful methodologies to automate and personalize the visualization process to better address the user's preferences, needs and interest. To address this main question, the work has been split into three sub-questions:

**RQ1: Which aspects of the visualizations have to be considered to automatically generate and recommend visualizations for the non expert user across various fields/domains?**

This research question aims to investigate how to structure the visualizations in a unified format that can be queried, reused by various technologies and extended with alternative semantics. In particular, we concentrate on encoding visual properties and functions of the visualizations in a standardized format which the visualization tools can interpret and transform in a graphical representation.

To answer this question, an extensive background (see Chapter 2) analysis has been conducted covering expert knowledge. The state of the art analysis presented in Chapter 2 has been applied to analyze and discuss relevant approaches with regard to their (i) domain independency, (ii) extensibility and (iii) format. Based on the insight gained from these reviews, a Visualization Analytics Vocabulary/ Visualization Ontology has been developed which describes the semantics of the visualizations in form of pragmatic simple facts that shall aid the sensible mapping from data to the visualizations. This vocabulary/ ontology has been used for the rule-based visual recommender which generates visualization recommendations



based just on data properties. The performance and usability of this prototype has been evaluated in a lab study [STV<sup>+</sup>14].

**RQ2: Can standard recommender system strategies be exploited to personalize the visualization recommendation in order to better address the user preferences/needs?**

Based on the insights gained from RQ1 a personalized visual recommender has been developed which exploits individual recommendation techniques (Collaborative filtering, content-based filtering and hybrid filtering) each considering different aspects of the user preferences to recommend personalized visualizations (see Chapter 5). The performance of each recommender technique has been evaluated performing independent offline evaluations. The results of the evaluations reveal which recommendation technique is more appropriate to recommend personalized visualizations for the user [MVT16].

**RQ3: Which methods for preference elicitation to use to define personalized visualizations that are closer to the user's needs and interests?**

This research question comprises the investigations of which information sources better describe (i) user's visual preferences and needs and (ii) the features of the visualizations to generate more accurate visual recommendations. To answer these questions, we analyzed the encoding power of different kind of information sources (tags, titles, questions&answers) in (i) encoding the user and (ii) encoding the visualizations applying information theoretic measures entropy, conditional entropy and mutual information [MVT17].

In summary, the first research question investigates the general approach to automatically define visual recommendations for the non-expert users from individual domains defining a unified convention to describe visualizations. The second research question discusses the possible benefits of existing recommendation strategies to be used in defining personalized visual recommendations. Finally, the third research question investigates the power of different kind of information sources in defining high quality visual recommendations.

### 1.3 Scientific Contribution

This thesis presents techniques and studies to design and evaluate a visual recommender. The aim of the thesis was to create a novel recommender system which, first, automatically defines a list of appropriate visualizations and next personalizes it considering user's preferences and needs. In the following, the scientific contribution of this thesis is outlined:

- *Visual Analytics (VA) Vocabulary*: This vocabulary presents a standard format to describe the visualizations. It comprises simple facts about the visualizations, such as visual properties, concepts and characteristics, that aid the sensible mapping from data to visualizations [MVT16]. The vocabulary has been inspired to a large extent by the pioneering work of Bertin who provided a theoretical foundation in information visualization and Mackinlay who developed a formal algebraic approach for the automatic design of graphical presentations. Following these general rules and principles in visual encoding, the VA vocabulary is designed to be applied in various applications across different domains. To support this, domain relevant formulations can be easily applied by just following the predefined syntax.

Furthermore, as a building block the VA Vocabulary is applied to define an OWL ontology of visualizations. Combined with a specialized algorithm (mapping algorithm) the ontology is used in a rule-based recommender system to automatically define the list of appropriate visualizations for a given dataset. Appropriate in this context means that the recommend visualizations are perceptually correct regarding visual encoding rules and guidelines.

In summary, the main contribution of the VA Vocabulary is to support the designer of a visual recommender, regardless from which domain she is, to define a system which automatically generates valid visualizations for the ordinary user.

- *Two-stage recommender system*: This novel recommender system aims to define a list of personalized visualizations which better reflects user's preferences and needs. It includes two stages, a rule-based stage which uses the visualization ontology and the mapping algorithm to define a list of appropriate visualizations for user's data and a personalization stage which personalizes the list according to user's preferences and needs. Basically, when using ontology based algorithms the list of possible visualizations results a permutations of visual configurations for each chart obtaining many perceptual incorrect ones. However, the presented mapping algorithm reduces this uncertainty caused by permutation. To achieve that, the algorithm applies visual patterns each describing a possible mapping for a concrete visualization.

Once the list with appropriate visualizations is defined, the system passes it to the next stage, the personalization stage, which applies different recommendation strategies, with regard to what is known about the user, to personalize the list according to the user's needs and preferences. In the personalization stage, three conventional recommendation techniques take place: Collaborative filtering (CF) build on the user ratings describing user's visual preferences, Content-based (CB) filtering build on the user tags (comments etc.) describing user's topic of interest and a hybrid recommender which combines CF and CB to a single technique. Typically, these techniques are used for product

recommendations. For this thesis, they have been adapted to recommend personalized visualizations too whereby their capability to generate high quality visual recommendations have been confirmed in various offline studies.

- *Concept to elicit the user preferences for visual recommender systems:* This thesis is based on the premise that the preference of a visualization is a personal matter. A crowd-sourced study was conducted to collect the empirical evidence supporting this assumption. In this study, participants were shown different visualizations, looking at the visualization they had to: 1) write tags (at least two), 2) write a title, 3) rate the visualizations using a multidimensional scale and 4) provide a question the visualization may answer. Beside confirming the varied visual preferences of people, the collected user preferences have been used as input for the integrated recommendation techniques to generate personalized visualizations.

The presented approach to collect visual preferences through a crowd-sourced study involving the general public and the gained insight from this study are, to the best of my knowledge, novel making them to one of the major contributions of this thesis.

- *Concept for measuring the efficiency of various information sources to be used for visual recommender systems:* In traditional recommendation approaches, the systems collect the user preferences in form of ratings or tags. However, in the context of visualizations the user can use alternative sources to communicate her preferences and interest. For instance, it is common for the user to pose a question that is answered with a visualization or to define a description for the visualization in form of a title. This thesis provides a concept describing how to investigate (i) the quality of tags, titles and question&answers in describing user's preferences and (ii) their capability to be used for a visual recommender. The results of this investigation should support the designer of visual recommenders by choosing the most efficient method (rating, tagging, question&answering etc.) for preference elicitation.

## 1.4 Research Papers

This thesis is supported on contributions described in eight peer-reviewed publications. They consist of two journal papers, four conference papers, one demo and one poster. The following provides a list of the publications and an overview about their scientific contribution.

**P1: Automated Visualization Support for Linked Research Data**

*Authors:* **Mutlu, B.**, Hoefler, P., Sabol, V., Tschinkel, G. and Granitzer, M.

*Contribution:* In this demo paper, the general concepts of the Visual Analytics (VA) Vocabulary is presented. The paper emphasizes the benefit of using a standard format to describe visualizations in order to facilitate the automated visualization process. The paper further introduces a mapping strategy that, together with the VA Vocabulary, provides the foundation of the rule-based recommender system, discussed in Chapter 3.

*Published:* Proceedings of the 9th International Conference on Semantic Systems.

*Year:* 2013

**P2: Suggesting Visualizations for Published Data**

*Authors:* **Mutlu, B.**, Hoefler, P., Veas, E., Sabol, V., Tschinkel, G., Stegmaier, F. and Granitzer, M.

*Contribution:* This publication details the rule-based recommender. Moreover, it presents a usage scenario including the tool chain for extracting statistical data from scientific publications, their semantic annotation and finally their automated graphical representation using the rule-based recommender. The tools for data extraction and their semantic annotation are provided by partners of the EU funded CODE (Commercially Empowered Linked Open Data Ecosystems in Research) project. Further, the publication contributes initial evaluation of the prototype which reveals that the rule-based recommender reduces the complexity of defining relevant visualizations for the ordinary user by creating them automatically.

*Published:* Proceedings of the 5th International Conference on Information Visualization Theory and Applications.

*Year:* 2014

**P3: Discovery and Visual Analysis of Linked Data for Humans**

*Authors:* Sabol, V., Tschinkel, G., Veas, E., Hoefler, P., **Mutlu, B.** and Granitzer, M.

*Contribution:* This journal paper extends the usage scenarios for the rule-based recommender. In particular, it contributes an extensive usability evaluation which confirms the results obtained for P2 that the ordinary user highly benefits from the rule based recommender. However, the results too provide the first evidence that the user preferences matter when defining meaningful visualizations for user's data.

*Published:* Proceedings of the 13th International Semantic Web Conference.

*Year:* 2014

**P4: Visual Recommendations for Scientific and Cultural Content**

*Authors:* Veas, E., **Mutlu, B.**, di Sciascio, C., Tschinkel, G. and Sabol, V.

*Contribution:* This paper is a follow up of P2 and presents the concept of defining appropriate visualizations for data from cultural and educational content to support their visual navigation and exploration. The concept presented here provides the foundation for the personalized visual recommender, discussed in Chapter 5.

*Published:* Proceedings of the 6th International Conference on Information Visualization Theory and Applications

*Year:* 2015

**P5: VizRec: A Two-Stage Recommender System for Personalized Visualizations**

*Authors:* **Mutlu, B.**, Veas, E., Trattner, C. and Sabol, V.

*Contribution:* This poster outlines the concept of the personalized visual recommender which results from the insights we gained in P1, P2 and P3. It introduces the applied recommendation strategy and presents the design and preliminary results of the crowd-sourced study performed to collect the user preferences. The results of the study reveal that preferences widely vary for visualizations generated automatically confirming our assumption that personalization is needed when defining visual recommendations.

*Published:* Proceedings of the 20th International Conference on Intelligent User Interfaces Companion.

*Year:* 2015

**P6: Towards a Recommender Engine for Personalized Visualizations**

*Authors:* **Mutlu, B.**, Veas, E., Trattner, C. and Sabol, V.

*Contribution:* This paper details the personalized visual recommender. It presents the design principles and architecture of the applied recommender technique. The paper further contributes the extensive analyzes of the collected user preferences and the performance of the recommender. The former analysis serves (i) to understand the variability in choice of preferred visualizations and (ii) to analyze assessments of quality of visualizations. Whereby the later assesses the quality of the defined recommendations with regard to their relevance to the user.

*Published:* Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization.

*Year:* 2015

**P7: VizRec: Recommending Personalized Visualizations**

*Authors:* Mutlu, B., Veas, E. and Trattner, C.

*Contribution:* This journal presents a recommendation pipeline including different recommendations techniques to define personalized visualizations. Each technique applies a specific kind of information source describing user's preferences and needs. In this context, the publication further studies which information lets the system anticipate user's choice of visualization for data analysis and how to represent such information and use it for a visual recommender.

*Published:* Journal ACM Transactions on Interactive Intelligent Systems- Special Issue on Human Interaction with Artificial Advice Giver.

*Year:* 2016

**P8: Tags, Titles or Q&As? Choosing Content Descriptors for Visual Recommender Systems**

*Authors:* Mutlu, B., Veas, E. and Trattner, C.

*Contribution:* This paper concentrates on characterizing the user preferences, in particular: (i) the source of information used to describe user's preferences and (ii) the methods to produce the most suitable recommendations thereby. It considers three sources corresponding to different aspects of interest: Tags, title and questions&answers. First, information-theoretic measures are applied to each source to measure it's the power in describing user's preferences and needs. Second, the practicability of each input type is evaluated with the visual recommender pipeline.

*Published:* Proceedings of the 28th ACM Conference on Hypertext and Social Media.

*Year:* 2017

P1, P2 and P3 present the rule-based recommender which automatically recommends the list with appropriate visualizations by just considering the properties of the data and the visualizations. P1 concentrates on the VA Vocabulary, whereby P2 too details the mapping algorithm. Furthermore, P2 and P3 provide the results of the lab studies that laid the foundation for the personalized visual recommender. P4 presents a usage scenario for the rule-based recommender and provides the preliminary ideas for a personalized visual recommender. P5 introduces the personalized recommender and present the preliminary results of the crowd-sourced study performed to collect the user preferences. P5 and P6 detail the recommendations pipeline describing each stage including an individual recommendation technique. P7 evaluates the power of different information sources in describing user's preferences and their capability to be used for visual recommender systems.

## 1.5 Structure of the Thesis

The rest of the thesis is organized as follow:

**Chapter 2** provides background information for this thesis. It outlines the key definitions, principles and concepts in Information Visualization, Recommender Systems and Semantic Web.

Moreover Chapter 2 presents previous research conducted in the area on modeling visualizations and automatically defining and recommending appropriate visualizations. It covers a summary of existing approaches, technologies and tools and outlines the past achievements in this field of research.

**Chapter 3** builds upon the need to a common format (i.e., domain-and technology independent) to describe visualizations. This chapter provides a vocabulary (Visual Analytics Vocabulary) to formulate expert knowledge about visualizations needed to create meaningful visualizations (**P1** [MHS<sup>+</sup>13], **P2** [MHT<sup>+</sup>14], **P4** [VMdS<sup>+</sup>15]). This vocabulary has been inspired to large extent by Bertin’s theory on information visualization [Ber83] and Mackinlay’s formal language to generate graphical presentations [Mac86].

Furthermore, Chapter 3 describes the concept of the rule-based recommender system which automatically recommends appropriate visualizations considering the properties of the data and semantics of the visualizations defined using the Visual Analytics Vocabulary (**P2** [MHT<sup>+</sup>14], **P7** [MVT16]). It analyzes the design and implementation of the fully functional prototype and its evaluation performed as a preliminary lab study involving eight IT experts (**P3** [STV<sup>+</sup>14]). Note that the rule-based recommender contributes to the **RQ1** together with the Visual Analytics Vocabulary (see Section 1.2).

**Chapter 4** details the crowd-sourced study conducted to collect user’s visual preferences in form of ratings, tags, titles and questions&answers. It presents the planning and procedure as well as the results of the study. It closes the section with studying the quality of the collected ratings in describing the user preferences and discusses the impact of the varied preferences of the user on recommending visualizations (**P5** [MVTS15b], **P6** [MVTS15a]).

**Chapter 5** builds upon the need of personalize the visualization recommendations. It introduces a novel visual recommender which generates personalized visualizations that better reflect user’s preferences and needs (**P7** [MVT16]). It provides detailed insights about the concept and architecture of the visual recommender which uses different strategies iteratively added into the recommendation pipeline. The first version was a result of a crowd-sourced study which revealed that preference of a visualization for data is a personal matter. Thereafter, each following iteration was guided by evaluation the performance of the previously applied technique that re-

sulted in more advanced one. This chapter shows the evolution of the strategy used to recommend visualizations.

The chapter about the personalized recommender ends with the evaluation of the applied recommendation strategies in defining appropriate visualizations. Each of this recommendation techniques considers different aspects of the user preferences revealed in form of ratings or tags. The evaluation studies the quality of the generated personalized visualizations when using ratings, tags or both applying a five-fold cross-validation. Note that the novel recommender system introduced in this chapter contributes to the **RQ2** (see Section 1.2).

**Chapter 6** extends the evaluation described in Chapter 5 and evaluates the capability of tags, titles and questions&answers(Q&As) in recommending personalized visualizations. The evaluation emphasizes the power of the individual input types in encoding user's preferences and the properties of the visualizations. To measure the encoding power, the chapter analyzes the use of information-theoretic measures (**P8** [MVT17]). The results presented in this chapter contributes to the **RQ3** (see Section 1.2).

**Chapter 7** concludes this thesis with answering the research questions, providing a self-reflection about the development and results of the thesis and providing an outlook on possible future work.

## 1.6 Summary

In this thesis we present methods to automate and personalize the visualization process to better address user's preferences, needs and interest. We focus thereby on addressing the following three research questions:

1. RQ1: Which aspects of the visualizations have to be considered to automatically generate and recommend visualizations for the non expert user across various fields of domains?
2. RQ2: Can standard recommender system strategies be exploited to personalize the visualization recommendation in order to better address the user preferences/needs?
3. RQ3: Which methods for preference elicitation to use to define personalized visualizations that are closer to user's needs and interest?

A comprehensive review of the existing expert knowledge about Information Visualization and Recommender Systems was necessary to address the above mentioned questions. This is presented in Chapter 2.



Chapter 2 presents also the previous research conducted in the area on modeling visualizations and automatically defining and recommending appropriate visualizations. This chapter discusses the past achievements in this field of research and outlines the contribution of this thesis.

Chapter 3 details our contributions in modeling visualizations and in automatically defining and recommending appropriate visualizations.

Chapter 4 presents the crowd-sourced study conducted to collect users' visual preferences in form of ratings, tags, titles and questions&answers. The collected ratings and tags are used as input for our personalized visual recommender which is introduced in Chapter 5.

Chapter 6 evaluates the capability of tags, titles and questions&answers in recommending personalized visualizations.

Finally, Chapter 7 concludes this thesis, provides a self-reflection about the thesis and an outlook on possible future work.

# Chapter 2

## Related Work

This chapter starts presenting the theory behind recommending appropriate visualizations, comprising the basic building blocks of visualizations and the methods used to create and recommend them automatically. Thereafter, the previous research in modeling- and recommending visualizations is provided that motivates this thesis.

### 2.1 Background: Recommending Visualizations

The goal of this thesis is to provide a system that can automatically recommend and create personalized visualizations. Yet, providing such a system typically requires expert knowledge about visualizations: it requires knowledge about how a visualization is build and which rules and guidelines to consider to create meaningful visualizations. Thus, Section 2.1.1 describes the main components of visualizations and provides a collection of visual encoding rules and perceptual guidelines that helps us to understand how a visualization is build and how one can come from data to its visual representation.

Having gained this knowledge, we need to represent it in a way it can be reasoned with. Section 2.1.2 describes Semantic Web components, we use to formalize the rules and principles for creating visualizations. The result of this formalization is a vocabulary (Visual Analytics Vocabulary, see Section 3.1) which we use to describe visualizations in a unified way. This vocabulary is further applied in a rule-based system to automatically define visual recommendations (see Section 3.2.1).

Finally, having a set of visualizations, it is necessary to investigate methods that help us to define which visualization is more appropriate for the users. Section 2.1.3 describes the basic concepts of recommender systems which we use to rank the visualizations according to users' visual preferences and needs. The ranked visualizations are then presented to the users as personalized recommendations.

Summarized, this section presents the theoretical underpinning on Information Visualization, Semantic Web and Recommender Systems, the fields that this dissertation touches to provide a personalized visual recommender.

## 2.1.1 Information Visualization

Visualizations are powerful tools to analyze and explore data. Providing the ability to uncover insights from large amount of data, visualizations facilitate the perception of patterns, trends and errors in data which might remain hidden by other means of analysis [War12]. As a result, people become able to eliminate guesswork and formalize hypothesis about their data or confirm the expected ones. However, defining valid visualizations is a complex task which typically requires expert knowledge. According to Ignatius and Senay this knowledge, among others, comprises the following categories: data characteristics, visualization vocabulary and graphical perception guidelines.<sup>1</sup> The following sections introduce Information Visualization as one of the main blocks of this thesis. Our hope is, this background information will help the reader of this thesis to better understand the context we set for our Visual Analytics Vocabulary and the rule-based recommender both introduced in Chapter 3.

### 2.1.1.1 Data Characteristics

Visualization is defined as the process of representing data. This makes data to one of the main components of visualizations. Yet, each visualization is used to either represent a certain type of data or an certain aspect of the same data. According to Munzner, there are four basic dataset types people can visualize: tables, networks, fields and geometry. Furthermore, data can be represented as cluster, sets and lists [Mun14]. These datasets are defined by different combinations of the five datatypes: items, attributes, links, position and grids.

In a nutshell, tables contain cells which are made up of items and attributes. In networks and trees, the items (=nodes) are collected with links. Fields are represented by grids which contain attributes and are organized by position. Finally, spatial geometry is characterized by position information.

Within the scope of this thesis, tables are the only dataset type used. Yet, **tables** are often used in information visualizations that are consist of rows and columns. In a simple flat table, each row shows an item of data and each column is an attribute of the dataset. An **attribute** is a specification that can be measured, observed or logged such as age, price, temperature. In this thesis, attribute are sometimes also called data dimensions (see Section 2.1.2.3). An **item** is a simple entity (cf., a row in a table) representing e.g., cities, people and shops. Yet, a combination of a row (=item) and a column (=attribute) represents a **cell** which contains a value of that pair. Figure 2.1 shows a flat table with its internal structure. Multidimensional tables, however, have a more complex structure: the data values are ordered in hierarchies.

Returning to the datatype **attribute**, there exist various attribute types a visual designer should consider [Mun14]. The most basic distinction are between **categor-**

---

<sup>1</sup><https://www.siggraph.org/education/materials/HyperVis/percept/visrules.htm>

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box		7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69		5 4-Not Specified	Small Pack	0.44	6/6/05
69		5 4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

**Figure 2.1:** In a flat table an item is represented by a row and an attribute by a column. Their intersections represent a cell which contains the value of the pairwise combination of a row and a column (from [Mun14], Figure 2.5).

**ical** and **ordered** attribute types whereby the ordered attributes are subdivided into **ordinal** and **quantitative**.

The data belongs to the type categorical, if they do not have an implicit ordering. Examples for categorical data are genres, city names and product types. The different categories indicate whether two objects are similar (female) or different (female vs. male).

Ordered data, however, have an implicit ordering and are further divided into ordinal and quantitative data. An example for ordinal data is the shirt size. Note that for ordinal data arithmetic operations are not possible. For instance, in case of shirt size the concept large minus medium would not be meaningful. However, it is known that medium is a size between small and large. Further examples for ordinal data are, music and movie charts, ranking in sport challenges and military ranks.

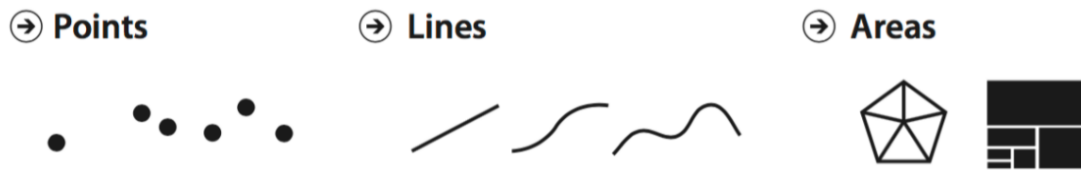
In contrast to the ordinal data, the quantitative data do support arithmetic operations. For instance, the quantity of 25 meter minus 10 meter is meaningful and the answer is 15 meter. Data describing height, weight, stock price and temperature as well as integers and real number are further examples of quantitative data.

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Small Pack	0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

**Figure 2.2:** Table with the attribute columns. The columns are color coded regarding to their type: quantitative, ordinal and categorical (from [Mun14], Figure 2.9).

For the sake of completeness, there is a further distinction of quantitative data into **interval** and **ratio data**. In case of interval data, the interval between each value is the same. For instance, when measuring the temperature, the interval between 100 and 90 is as same as the interval between 70 and 60. Ratio data are basically interval data but with a true zero point. For instance, the difference between 2 and 3 is the same as between 4 and 5 as well as 4 grams is twice a weight of 2 grams. Other examples for ratio data are, age and length. Nevertheless, as suggested by Tamara Munzner [Mun14], the categorization of data into interval and ratio is not meaningful when defining visual encoding. Thus, this thesis collapses interval and ratio data into the single category quantitative data and concentrates on categorical, ordinal and quantitative attribute types.

For a better understanding, Figure 2.2 shows a simple table with the attribute columns where the columns are color coded according to their type.



**Figure 2.3:** There exist four types of marks: point, line, area and volume (from [Mun14], Figure 5.2).

### 2.1.1.2 Visualization Vocabulary

The aim of the Visualization Vocabulary is to define the fundamental building blocks of Information Visualization.<sup>2</sup> **Marks** are the most basic building blocks that encode links or items. Basically, marks are geometric objects classified according to the number of spatial dimensions they require. There are four types of marks: a zero dimensional mark is a point, a one dimensional mark is a line, a two dimensional mark is an area and a three dimensional mark is a volume [Mun14]. Figure 2.3 illustrates the individual mark types.

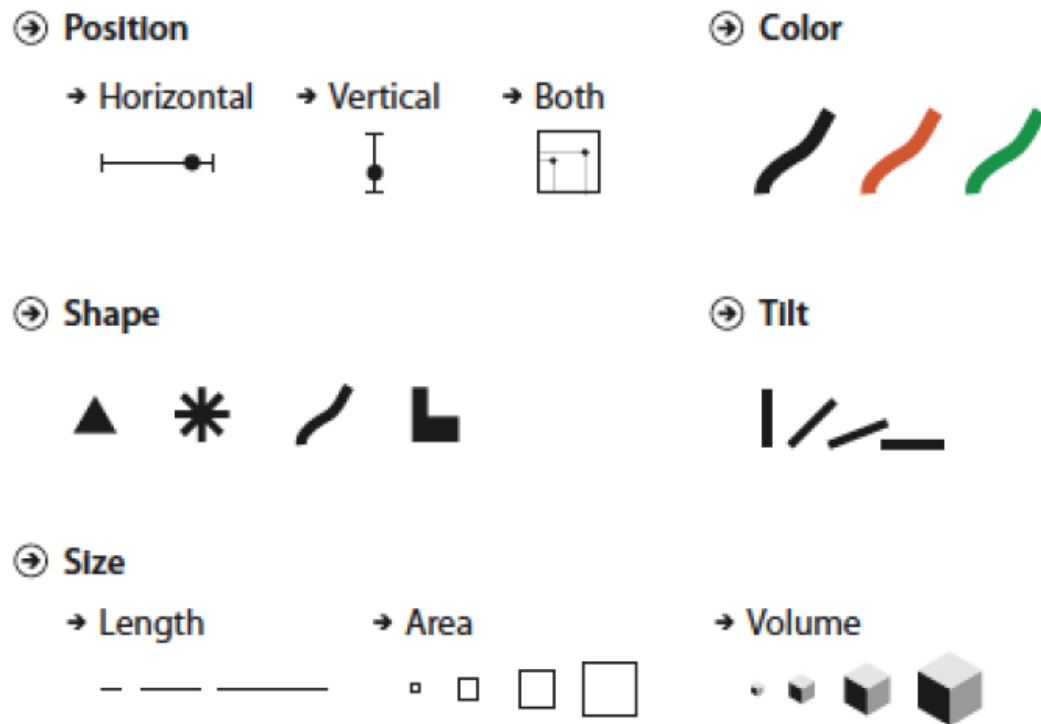
The appearance of the marks is controlled by so called **visual channels** [Mun14] which encode the properties of the marks. The classification of the visual channels were coined by Jaques Bertin [Ber83] who originally called them visual variables. In the visualization literature, there exist many other synonyms for visual channels. However, within the scope of this thesis we use the term visual channels to describe the symbols applied to encode the marks. Bertin’s list of visual channels comprised position, size, shape, value, color orientation and texture. This list has been extended with additional channels by Jock D. Mackinlay [Mac86]: length, angle, slope, area, volume, density, color saturation, color hue, connection and containment.

Figure 2.4 shows some of the mostly used visual channels. In a nutshell, there are four position channels: aligned planar position, unaligned planar position, depth (three-dimensional position) and spatial region. Other visual channels are represented by color with its three aspects: hue, saturation and luminance. There exist three size channels: length in one-dimensional size, area in two-dimensional size and volume in three-dimensional size. Further visual channels are tilt (also called angle) and shape.

With regard to human perceptual system, there are two types of visual channels: **identity channels** providing information about *what* or *where* something is and **magnitude channels** providing information about *how much* of something there is [Mun14]. For instance, shape and color channel hue are identity channels. When used on a spatial region mark, they can provide the information about *where* the region is.

The examples for magnitude channels are, length, color channel luminance, angle,

<sup>2</sup><https://www.siggraph.org/education/materials/HyperVis/concepts/marks.htm>



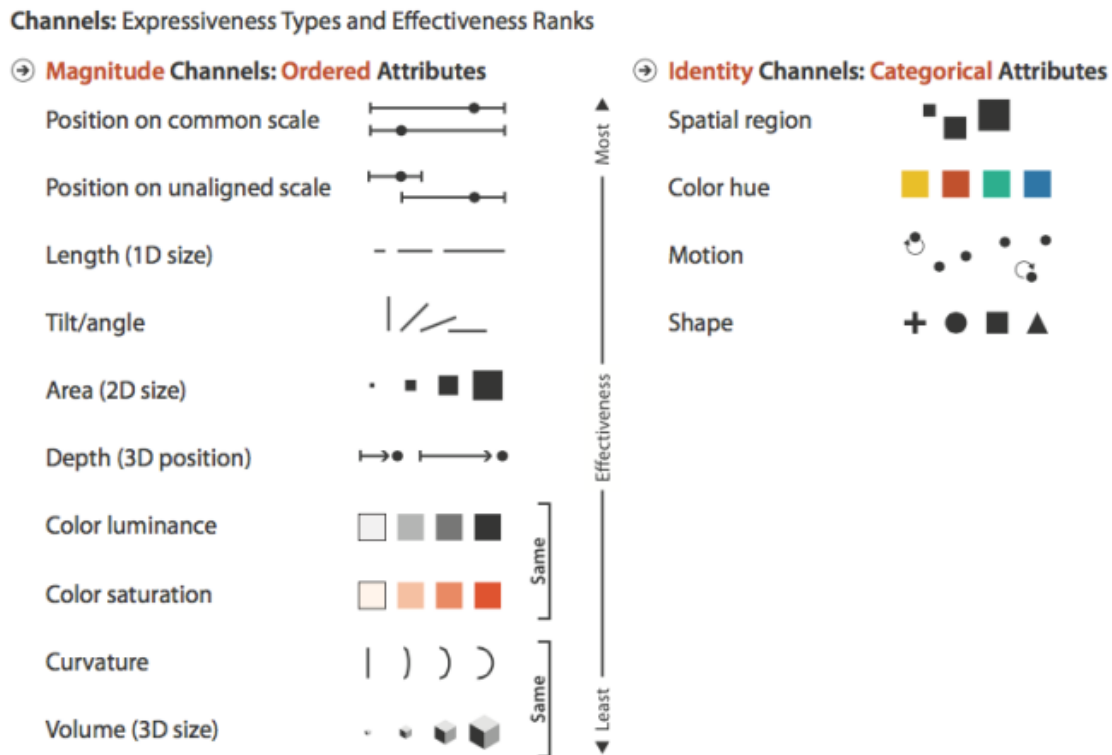
**Figure 2.4:** Visual channels. There are four position channels: aligned planar position, unaligned planar position, depth and spatial region. Other visual channels are represented by color with its three aspects: hue, saturation and luminance. There exist three size channels: length in one-dimensional size, area in two-dimensional size and volume in three-dimensional size. Further visual channels are tilt (also called angle) and shape (from [Mun14], Figure 5.3).

area and volume. For instance, when using the magnitude channel length, one can ask *how much* longer a certain line is than another. Similar to this, when using luminance, one can address the question on *how much* darker one mark is than another.

### 2.1.1.3 Graphical Perception Guidelines

Graphical perception is the ability of human’s visual system to accurately decode the quantitative and categorical information which have been previously encoded using visual channels and marks [CM85, Mac86]. Yet, it depends on the properties of the used visualizations how effectively people perceive and process the encoded information.

Over the years, a great deal of researchers in cartography, statistics and computer sciences have formalized perceptual guidelines which a visual designer should consider when defining effective visual representations [Ber83, Mac86, HB10a]. The



**Figure 2.5:** Channel ranking according to their effectiveness in encoding data attributes. Note, magnitude channels should be used for ordered data and identity channels for the categorical data (from [Mun14], Figure 5.6).

guidelines comprise the criteria for the visual encoding, perceptually-motivated rankings and characteristics of the visual channels.

Mackinlay developed a formal visual encoding language to generate graphical presentations for relational information [Mac86]. He defined two principles a visual designer should follow when using visual channels: **expressiveness** and **effectiveness**.

“The expressiveness principle dictates that the visual encoding should express all of, and only, the information in the dataset attributes” [Mun14, Mac86]. What this means is, ordered data should be presented in a way that human perceptual system recognizes them as ordered. In contrast, unordered data should not be presented in a way that people sense them as ordered data. As a result, the identity channels are defined as appropriated candidates for the categorical attributes that have no specific order. In contrast, the magnitude channels represent the appropriate candidates for ordered attributes, both ordinal and quantitative.

“The effectiveness principle dictates that the most important attributes should be encoded with the most effective channels to be most noticeable” [Mun14].

Summarized, expressiveness depends on the syntax and semantic of the graphical



language whereby effectiveness also depends on the capabilities of the perceiver. Based on that, Mackinlay defined a ranking of visual channels for perceptual tasks that associated with the interpretation of graphical representation.

Figure 2.5 illustrates Mackinlay’s effectiveness ranking for the different channel types [Mac86]. In this figure, the most effective channels are presented on the top and the less effective channels on the bottom. In a nutshell, the most effective magnitude channel is aligned spatial position followed by unaligned spatial position, length (1D size), angle, area (2D size) and finally by depth (3D position). However, the most effective identity channels are spatial region, color hue, motion and finally shape.

#### 2.1.1.4 Summary

This section introduces Information Visualization as as one of the main blocks of this thesis. It identifies the main components of visualizations needed to understand how the visualizations are created, and introduced the visual encoding rules and perceptual guidelines important to create meaningful visualizations. Using the information provided here, we are able to (i) define a common format to describe visualizations, the Visual Analytics Vocabulary (see Section 3.1), and (ii) develop a rule-based recommender which uses this format to automatically recommend appropriate visualizations for a given dataset (see Section 3.2.1). Another purpose of this section is to provide enough information for the reader of this thesis who are unfamiliar with the theory of Information Visualization to better understand the subject, methodology and findings presented in Chapter 3.

The next section introduces the research topic Semantic Web with its components we used to define our Visual Analytics Vocabulary. Basically, we use these components to formalize rules for creating visualizations and to apply semantics to visualizations. These components are introduced in Section 2.1.2.3. The rules, however, are defined following the expert knowledge presented in Section 2.1.1.

Yet, we made use of Semantic Web components not only for describing visualizations but also some of the data we used as input for our rule-based visual recommender. These components are introduced in Section 2.1.2.3.

### 2.1.2 Semantic Web

The idea of Semantic Web has been proposed by Tim Berners Lee. Lee describes Semantic Web as “the structure to the meaningful content of Web pages, creating an environment where software agents, roaming from page to page, can readily carry out sophisticated tasks for users” [BLHL01]. These tasks comprise data processing and interpretation. Yet, Semantic Web is not meant to replace the current World Wide Web. It should rather extend it so that advanced knowledge management becomes possible on the network of hyperlinked human-readable web pages. To achieve that, Semantic Web makes use of so called machine-readable metadata which provides

<b>Subject</b>	<b>Predicate</b>	<b>Object</b>
Shakespeare	wrote	King Lear
Shakespeare	wrote	Macbeth
Anne Hathaway	married	Shakespeare
Shakespeare	livedIn	Stratford
Stratford	isIn	England
Macbeth	setIn	Scotland
England	partOf	UK
Scotland	partOf	UK

**Figure 2.6:** Sample triples: a list of plays that Shakespeare wrote or a list of parts of the United Kingdom. From [AH11], Table 3.3.

information about each page and their relation to each other. This, in turn, makes it possible for the software agents to readily access the websites and interpret or even process them for a further use.

The following sections introduce a selection of techniques behind Semantic Web that has been used within this thesis.

### 2.1.2.1 Resource Description Framework (RDF) & RDF Schema (RDFS)

Resource Description Framework (RDF) is a data model for describing resources and their relationship to each other. Each statement of a relationship is defined in form of **subject-predicate-object** expressions, called triples [KS15a]. Hereby, a **subject** represents the resource, the **predicate** the relationship between **subject** and **object** and the **object** the actual value. Figure 2.6 shows a simple example of a triple. Each element of a triple is represented with a URI (Uniform Resource Identifier) serving as a global identification for the resource that is common across the WWW [AH11]. An URI might be, for instance, a URL (Uniform Resource Locator).

One concern regarding to RDF is, it cannot be used to describe ontology relations and constraints. To tackle this issue, the RDF Schema (RDFS) has been defined which serves as a semantic extension to the RDF vocabulary.<sup>3</sup> RDFS specification provides a machine-understandable formalism used to describe classes, their properties, relationship and constraints on the allowed combinations of classes [Maj08]. Thus, it defines the building blocks of an ontology.

In an nutshell, the classes in RDFS vocabularies are:

- **rdfs:Resource** describes all resources
- **rdfs:Property** the class of properties

<sup>3</sup><http://www.w3.org/TR/rdf-schema/>

- **rdfs:Class** declares a resource as a class
- **rdfs:Datatype** states the class of datatypes

Properties, however, are instances of `rdf:Property`, describing the relation between resources:

- **rdf:type** is a property used to highlight that a resource is an instance of a class
- **rdf:subClassOf** construct to build the hierarchy
- **rdfs:subPropertyOf** declares that all resources which are related by property are related by another
- **rdfs:label** used to state the name of the resource
- **rdfs:comment** used to describe the resource
- **rdfs:isDefinedBy** used to refer to an RDF vocabulary in which a resource is described

The RDFS specification also introduced constraints for the properties which are:

- **rdfs:domain** declares the class of the *subject*
- **rdfs:range** declares the class or datatype of the *object*

### 2.1.2.2 The W3C Web Ontology Language (OWL)

Ontologies are used to describe and represent domain specific knowledge (e.g., medicine) which can be then shared by applications, people and databases. Using ontologies the designer specify the classes (things), relationships between classes and properties of classes. Summarized, the aim of OWL is to define a schema the designer can use to describe and process the information content within ontologies that make them processable both by machines and humans [KS15b].

### 2.1.2.3 The RDF Data Cube Vocabulary

The RDF Data Cube Vocabulary is a W3C Standard defined to translate multi-dimensional data, such as statistics, into a format in which these data can be easily linked and combined with related information. This is possible, because the Data Cube Vocabulary follows the W3C RDF (Resource Description Framework) standard and the principles of Linked Data. For exchanging and sharing statistical data and metadata among organizations, the vocabulary follows the SDMX (Statistical Data and Metadata eXchange) ISO standard [CRT13]. Since the model is very general, it can be applied to model any kind of statistical data. This is shown in

**Table 2.1:** Life expectancy broken down by region (unitary authority), age and time. From [CRT13], Table 1.

	2004-2006		2005-2007		2006-2008	
	Male	Female	Male	Female	Male	Female
Newport	76.7	80.7	77.1	80.9	77.0	81.5
Cardiff	78.7	83.3	78.6	83.7	78.7	83.4
Monmouthshire	76.6	81.3	76.5	81.5	76.6	81.7
Merthyr Tydfil	75.5	79.1	75.5	79.4	74.9	79.6

Section 3.2.2 where we illustrate how this vocabulary is applied to model statistical data extracted from scientific publications to automatically visualize and analyze them using our-rule based recommender.

A dataset represented as a RDF Data Cube is basically a collection of observations representing some statistics. The observations consist of so called dimensions, measures and attributes. The dimensions identify the observation, the measures are related to concrete values and attributes add semantics to them. The vocabulary represents these components as RDF properties:  $qb : DimensionProperty$ ,  $qb : AttributeProperty$  and  $qb : MeasureProperty$  derived from generic observation class  $qb : ComponentProperty$  [CRT13]. For instance, the Table 2.1 shows the life expectancy from state Wales<sup>4</sup> broken down by region, age and time. Looking at the table, we recognize three dimensions: the time period, region and gender. This table has one measure which is the life expectancy so that each observation represents the life expectancy for each region. When transformed in a Data Cube, there will be an attribute which defines the units (years) of the measured values. Yet, after the transformation, the table looks like as illustrated in Listing 2.1.

Looking at the Listing 2.1, we observe that first the components (dimensions, measures) of the Data Cube have to be defined. There are suitable predefined concepts in the SMDX-COG for period ( $sdmx - dimension : refPeriod$ ), area ( $sdmx - dimension : refArea$ ) and sex ( $sdmx - dimension : sex$ ) which can be reused to represent these dimensions. Note, the possible values of the components are defined using the  $rdfs : range$  property. However, when referring to a  $qb : MeasureProperty$ , it is common to use  $sdmx - measure : obsValue$ .

As next, the data structure of the cube has to be defined including all dimensions and measures (see Listing 2.2). If required, it is possible to define an order for the components by giving an integer value for  $qb : order$ . Yet, when using  $qb : MeasureProperty$  it is possible to attach only a single attribute to each observation which will describe only one measured value [KRK<sup>+</sup>15]. To tackle this issue, the  $qb : componentAttachment$  property can be used which attaches one attribute to each  $qb : MeasureProperty$  and regards the whole data set ( $qb : DataSet$ ) [KRK<sup>+</sup>15]. Finally, each individual observation looks like as shown in Listing 2.3.

<sup>4</sup>Wales is a country that is part of the United Kingdom.

**Listing 2.1:** Dimensions and Measures of the RDF Data Cube. From [CRT13], Example 1, 2, 3.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix eg: <http://example.org/ns#> .
@prefix qb: <http://purl.org/linked-data/cube#> .
```

```
eg:refPeriod a rdf:Property, qb:DimensionProperty;
  rdfs:label "reference_period"@en;
  rdfs:subPropertyOf sdmx-dimension:refPeriod;
  rdfs:range interval:Interval;
  qb:concept sdmx-concept:refPeriod .
```

```
eg:refArea a rdf:Property, qb:DimensionProperty;
  rdfs:label "reference_area"@en;
  rdfs:subPropertyOf sdmx-dimension:refArea;
  rdfs:range admingeo:UnitaryAuthority;
  qb:concept sdmx-concept:refArea .
```

```
eg:lifeExpectancy a rdf:Property, qb:MeasureProperty;
  rdfs:label "life_expectancy"@en;
  rdfs:subPropertyOf sdmx-measure:obsValue;
  rdfs:range xsd:decimal .
```

**Listing 2.2:** Data Structure of the RDF Data Cube. From [CRT13], Example 4.

```
eg:dsd-le a qb:DataStructureDefinition;
  # The dimensions
  qb:component [ qb:dimension eg:refArea; qb:order 1 ];
  qb:component [ qb:dimension eg:refPeriod; qb:order 2 ];
  qb:component [ qb:dimension sdmx-dimension:sex; qb:order 3 ];
  # The measure(s)
  qb:component [ qb:measure eg:lifeExpectancy];
  # The attributes
  qb:component [ qb:attribute sdmx-attribute:unitMeasure;
    qb:componentRequired "true"^^xsd:boolean;
    qb:componentAttachment qb:DataSet; ] .
```

**Listing 2.3:** Individual observations RDF Data Cube. From [CRT13], Example 9.

```

eg:dataset-le1 a qb:DataSet;
  rdfs:label "Life_expectancy"@en;
  rdfs:comment "Life_expectancy_within_Welsh_Unitary_authorities_-_extracted_from_Stats_Wales"@en;
  qb:structure eg:dsd-le ;
  .

eg:o1 a qb:Observation;
  qb:dataSet eg:dataset-le1 ;
  eg:refArea ex-geo:newport_00pr ;
  eg:refPeriod <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y
    > ;
  sdmx-dimension:sex sdmx-code:sex-M ;
  sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Year> ;
  eg:lifeExpectancy 76.7 ;
  .

eg:o2 a qb:Observation;
  qb:dataSet eg:dataset-le1 ;
  eg:refArea ex-geo:cardiff_00pt ;
  eg:refPeriod <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y
    > ;
  sdmx-dimension:sex sdmx-code:sex-M ;
  sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Year> ;
  eg:lifeExpectancy 78.7 ;
  .

...

```

#### 2.1.2.4 Summary

This section introduces the Semantic Web components we used to (i) define a common format to describe visualizations following the rules of Information Visualizations (see Section 3.1), and (ii) to model the input data for our rule based recommender (see Section 3.2.2).

The next sections introduce the research topic Recommender Systems. We thereby focus on two most conventional recommender techniques: collaborative filtering and content based filtering. We identify the key definitions, provide an overview about the used algorithms and talk about the challenges each technique has. We use the information and knowledge provided here for creating a visual recommender which automatically recommends personalized visualizations to the user (see Chapter 5). Yet, the following sections not only define the frame for our visual recommender but support the reader of this thesis to become familiar with terms, methods and techniques we mention in Chapter 5.

### 2.1.3 Recommender Systems

Recommender Systems are powerful software tools which support us by finding relevant and interesting items within a huge information space. We cross them in our daily digital life, in commercial e-commerce sites, in entertainment portals, social media, news portals and so on. Yet, there are several ways to generate recommendations. One way to define recommendations is using statistical data from the

community (=what is popular among the majority must also be interesting to me) without considering the preferences of the single users. In contrast, personalized recommendations take into account user's past and/or current preferences, needs and interest with the aim to define recommendations that are closer to what the user usually prefers.

The basis of defining recommendations is the information filtering [KC16]. The mostly used techniques therefore are collaborative filtering and content based filtering. These individual filtering techniques are introduced in the following sections. Note, these sections are mainly based on the publication P7 [MVT16].

### 2.1.3.1 Collaborative Filtering

Collaborative filtering (CF) is the most popular and successful filtering technique [SFHS07, SK09] which uses a collection of the user preferences to generate personalized visualizations. The preferences are nothing else but collections of either explicit ratings, for instance on a 1 to 7 scale, provided by the user in the past or implicit ratings which are automatically inferred from user's behavior.

CF uses this repository of known preferences of a group of users to define predictions of unknown preferences for other users. Hence, the basic idea behind it is: the users that had similar tastes or behaviors (e.g., reading, watching, buying etc.) in the past will have similar tastes or behaviors in the future.

The CF algorithms represent the ratings given by the users to items as an  $m \times n$  Matrix  $A$ . Each entry  $a_{i,j}$  in the Matrix  $A$  represents ratings of the  $i$ th user on the  $j$ th item. To generate the top- $n$  recommendations for the active user  $u$ , it is necessary to calculate the  $k$  most similar users or items (nearest neighbors) to the user  $u$ . There are two different CF approaches to obtain the nearest neighbors, namely (1) memory-based (user-based) CF and (2) model based (item-based) CF.

A popular similarity measure in CF is the *Pearson correlation* which measures the strength of the linear association between two variables and defines the direction (positive +1 or negative -1) of the association. For the user-based CF, the Pearson correlation between the user  $u$  and the user  $v$  is:

$$sim(u, v) = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}} \quad (2.1)$$

where  $I$  is the set of items rated by the users  $u$  and  $v$  and  $\bar{r}_u$  is the average rating of the active user  $u$ .

For the item based CF, the *Pearson correlation* is:

$$sim(i, j) = \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U} (r_{u,i} - \bar{r}_i)^2} \sqrt{\sum_{u \in U} (r_{u,j} - \bar{r}_j)^2}} \quad (2.2)$$

where  $U$  is the set of users who rated both items  $i$  and  $j$ ,  $r_{u,i}$  is the rating of the user  $u$  on the item  $i$  and  $\bar{r}_i$  is the average rating of the  $i$ th item by those users.

To make a prediction whether the active user  $u$  would like the unseen item  $i$ , the user-based CF adds/subtracts the neighbors bias from the active users average ratings and use this as a prediction for the item  $i$ . The algorithm applies the following equation 2.3:

$$pred_{cf}(u, i) = \bar{r}_u + \frac{\sum_{v \in N} sim(u, v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in N} sim(u, v)} \quad (2.3)$$

where  $\bar{r}_u$  and  $\bar{r}_v$  are the average ratings of the user  $u$  and  $v$ .

In contrast, the item-based algorithms make a prediction by averaging the ratings of similar items rated by the active user  $u$ :

$$pred_{cf}(u, j) = \frac{\sum_{i \in ratedItem(u)} sim(i, j) * r_{u,i}}{\sum_{i \in ratedItem(u)} sim(i, j)} \quad (2.4)$$

where  $j$  is represents the unseen item.

Summarized, the prediction will be defined using (1) the average ratings made by the user  $u$  and the user  $v$  and using (2) the average ratings of similar items rated by the active user  $u$ . In both cases the prediction is a numerical value and is within the same scale like user's ratings, i.e., from 1 to 7. After the predictions are calculated, the items are sorted in decreasing order based on their prediction value and placed in so called top-n list and recommended to the active user.

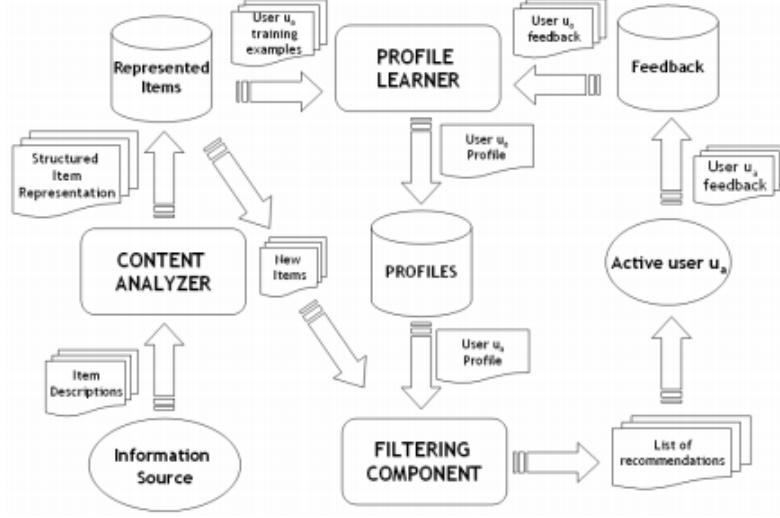
### 2.1.3.2 Content-based Filtering

Content-based (CB) recommender systems define recommendations based on the correlation between the content of the items and the user's profile [LdGS11]. In CB recommender systems the item content is represented with a set of extracted terms or features. In contrast, the personal comments and tags a user gives to items define her profile since the user-provided comments/tags are assumed to describe the user's taste, needs and interest, her topic of interest respectively.

Figure 2.7 illustrates the high level architecture of a content-based recommender system. In a nutshell, the CB recommender algorithm passes three stages: content analyzer, profile learner and filtering component [RRSK10]. The content analyzer stage addresses the task of extracting relevant information from the items and representing them as content descriptors. These descriptors are further used to profile the item. The next stage, profile learner, collects information describing user's preferences and interests that are used to construct the user profile. Finally, the filtering component uses the item and the user profiles to generate recommendations by analyzing if there are any correlations between them.

Usually, CB recommender systems uses keyword based *Vector Space Model (VSM)* together with basic *TF-IDF weighting* to determine the correlation between the items and the users. Transported in VSM, each item is represented as a vector of term weights where each weight indicates the degree of association between the item and the term. Similar to this, the user profiles can be represented by profile vectors.





**Figure 2.7:** High level architecture of Content-based recommender. From [RRSK10], Figure 3.1.

To clarify this, assume  $I = \{i_1, i_2, i_3, \dots, i_N\}$  represents a set of items and  $T = \{t_1, t_2, t_3, \dots, t_n\}$  a set of terms describing  $I$ . Each item  $i_k$  is represented as a vector in a  $n$ -dimensional vector space, i.e.,  $i_k = w_{1,k}, w_{2,k}, w_{3,k}, \dots, w_{n,k}$ , where  $w_{j,k}$  denotes the weight for the term  $t_j$  in document  $i_k$ , i.e.:

$$w_{j,k} = tf - idf_{t_j, i_k} = tf_{t_j, i_k} \times idf_t = tf_{t_j, i_k} \times \left[ \log_e \left( \frac{N}{df_t + 1} \right) + 1 \right] \quad (2.5)$$

The same weighting schema is used for the user profile. Having defined the weights, the recommender algorithm applies the *cosine similarity measure* to estimate the similarity between the user- and the item profiles which is:

$$sim(i_k, i_l) = \frac{\sum_j \mathbf{w}_{j,k} \mathbf{w}_{j,l}}{\sqrt{\sum_j (\mathbf{w}_{j,k})^2} \sqrt{\sum_j (\mathbf{w}_{j,l})^2}} \quad (2.6)$$

where  $i_l$  denotes the tag collection of the current user. The result of this measure is a cosine value of the angle between two vectors. Finally, this value is used as score to rank the relevant items regarding to their similarity to the user needs and interests following the Equation:

$$pred_{cb}(i_k, i_l) = \sum_{i_k, i_l \in I} sim(i_k, i_l) \quad (2.7)$$

### 2.1.3.3 Hybrid Filtering

The presented recommender techniques, collaborative and content-based filtering, have both advantages and shortcomings. The advantages of the CF include the content independency of the items being recommended, the low cost for knowledge acquisition and maintenance (no knowledge engineering is required) and the ease of use. However, this recommendation techniques suffer from so-called cold start problem [SPUP02]. The term *cold start* in the context of recommender systems generally characterizes the situations where the user have not yet provided her feedback to the system or when there is a new item transferred into the system, i.e., no past information is available. Even when there is some feedback provided, the collaborative filtering mechanisms in particular sometimes fail to provide the results, since they become unable to find the corresponding user with similar tastes. This is often referred as *data sparsity* problem [GSK<sup>+</sup>99] implying the collaborative filtering algorithm might be unable to form recommendations due to lack of information on the user or the item. In contrast, CB-RS do not require a direct user involvement in terms of e.g., providing ratings. Furthermore, these recommender techniques are capable of recommending items not yet rated by any user. However, the recommendations generated using content-based recommender systems can be too general since the systems might capture only a certain aspect of the content. In this case the user might be recommended items similar to those she already rated or tagged without considering her interest changing over time. One obvious solution for these problems is to combine different recommender systems to a hybrid recommender which uses the strength of all available recommender techniques.

There are different methods for a hybrid design [Bur02, JZFF10], including (i) weighted hybrids, (ii) switching hybrids, (iii) mixed hybrids, (iv) feature combination hybrids, (v) cascade hybrids, (vi) feature augmentation hybrids and (vii) meta-level hybrids.

### 2.1.3.4 Summary

In this section we introduce the recommendation techniques, we use within the course of this thesis to define a visual recommender that can automatically recommend personalized visualizations (see Chapter 5). We provide an overview about the used algorithms, identify the key terms and present the challenges by generating personalized recommendations.

## 2.1.4 Summary

This section provides a frame of topics addressed in this thesis, derived from Information Visualization, Recommender Systems and Semantic Web. In particular, it clarifies the main components, methods and techniques used to define meaningful visualizations, provides a detailed overview about the most popular Recommender

System strategies and introduces Semantic Web technologies used within the course of this thesis.

The next section 2.2 discusses the relevant work conducted in the area on visualization modeling and automatically recommending appropriate visualizations for a given dataset. It summarizes the existing methods, tools and technologies and outlines the past achievements in this field of research.

## **2.2 State-of-the-art: Recommending Visualizations**

Recommending visualizations is a relatively new area of research and only few efforts have been made so far to tackle the challenge. The relevant research is build on the visual encoding rules and perceptual guidelines and ignore the fact that the choice of visual representation involves as much user preferences and needs [MVT16]. This chapter provides an overview on previous research upon which the contribution of this thesis builds. Moreover, it specifies the research areas the thesis touches: modeling visual representations, recommending appropriate visualizations and recommending personalized visualizations.

Modeling visual representations aims to make visualizations available as highly structured representations that can be queried, integrated and extended [DFVR10]. It defines a common vocabulary for the experts to describe visualizations and at the same time it enables the ordinary users to generate visualizations automatically. The basis of such models are collected and stored expert knowledge about visualizations and, if required, the domain. One method for modeling visualizations is using an ontology language that has a well-defined syntax, well-defined semantics, efficient reasoning support [KVWW07] and a human and machine-readable formal specification. Typically, the models that organize the information about visualizations are used in rule-based systems which automatically generate and recommend appropriate visualizations based on correspondence of data attributes and visualization properties queried from the corresponding model.

To date, only a few approaches attempted to automatically generate visual representations starting just from the data [MHT<sup>+</sup>14, NRB<sup>+</sup>13], albeit with limited success. Despite their usefulness, these approaches ignore the fact that the choice of visual representation involves as much user preferences and needs: popular visualization tools require several human choices to tailor the end result to the user's preferences. Beyond visualization, recommender systems address the personalization issue based on knowledge about interests and previous choices of users. A number of questions arise to this respect. Which aspects of the visualization and underlying data are important for the user? How should these aspects be captured? Which strategies should be used to recommend visualizations based on them?

The chapter starts with summarizing the research performed on modeling visualizations with the focus on visualization ontologies. Thereafter, current trends in

visualization recommendation with focus on rule-based system are introduced, followed by behavior-based, semantic-based and user preference oriented approaches. With this frame, the previous works in visualizations recommendation are analyzed that motivate this thesis.

## 2.2.1 Modeling Visual Representations

Visualization can be considered from information theory as visually coding and communicating information [CJ10]. Bertin's work on semiology offers a systematic study of visual representations [Ber83]. It defines and characterizes visual variables that compose visualizations. Carpendale analyzes visual variables for computational information visualization [Car03]. Building on semiology, Mackinlay developed a formal language to generate graphical presentations for relational information and defined *expressiveness*: whether a graphical language can express the desired information and *effectiveness*: whether the graphical language exploits the capabilities of the output medium and the human visual system [Mac86]. Card and Mackinlay categorized data in terms of its attributes (e.g., nominal ordinal, quantitative). Engelhardt systematically analyzed syntactic structure and information type in graphic representations [vE02]. The structure and design of any graphical representation have a perceptual connotation with cognitive implications [War12]. These contributions build our understanding of the visual encoding principles that help us define visualizations [MVT16]. Yet, this rich formalization (visual encoding rules and perceptual guidelines) has to be represented in a common format to bridge the gap between formal description and visualization services to automatically create appropriate visualizations.

There exist different strategies to define a common format (model) for visualizations with respect to visual encoding rules and perceptual guidelines. A broad body of research formulates a taxonomy, for instance, using a formal language such as XML (Extensible Markup Language) or an ontology using, for instance, OWL. The following sections present a literature review on methods and technologies for modeling visualizations using one of the above mentioned strategies. The literature review is restricted to the years between 2002 and 2015. However, we do not claim completeness for all publications available in this area.

### 2.2.1.1 Visualization Modeling in XML

Wollowski [Wol15] provides an approach for searching information plotted on visualizations by extracting and representing them in a XML file conforming to a XML Schema. An example thereof is provided by the Figure 2.8. In a nutshell, the proposed approach suggest to specify every single chart in a XML file with a *type* and *title* description. The former identifies the visualization whereby the later the kind of information the visualization presents. The information encapsulated in the *plotarea* element specify the visual channels of the visualization with the data

```

<?xml version="1.0" ?>
<chart type="bar">
  <title>4-Year GPA by Gender</title>
  <legend />
  <plotarea>
    <axis variable="x1">Year</axis>
    <axis variable="x2" min="0" max="4.0"
      step="0.5">GPA</axis>
  </plotarea>
  <dataset label="Men">
    <datapoint x1="2000" x2="2.6" />
    <datapoint x1="2001" x2="2.8" />
    <datapoint x1="2002" x2="2.9" />
    <datapoint x1="2003" x2="3.0" />
  </dataset>
  <dataset label="Women">
    <datapoint x1="2000" x2="2.5" />
    <datapoint x1="2001" x2="2.8" />
    <datapoint x1="2002" x2="3.0" />
    <datapoint x1="2003" x2="3.1" />
  </dataset>
</chart>

```

**Figure 2.8:** XML File presenting a visualization including the data it shows. From [Wol15], Figure 3.

attributes mapped to them. Each mapped data value, however, is represented in a *datapoint* element. In total, this specification has been applied to describe, store and present four types of visualizations: bar chart, line chart and pie chart respectively.

Ferres et. al developed *iGraph*, a system which allows blind and visually-impaired users gain access to graphical representations of time-series data [FPL<sup>+</sup>06]. To achieve this, the system benefits from a reasoning engine and a Natural Language Interface (NLI) which provides a query mechanism to interact with the graph. The system consists of three main sub-systems: the P-, C-. and L-Systems. The P-System is responsible for acquiring information about the graph and representing them as a XML File. The C-System provides a set of rules for querying the graph data. Finally, the L-System loads the graph data and provides a dialogue-based NLI to enable the user interact with the loaded information. Noteworthy, the XML File produced in the P-System encodes the necessary elements and attributes of the graph based on the XML Schema defined by Wollowski [Wol15] shown in Figure 2.8.

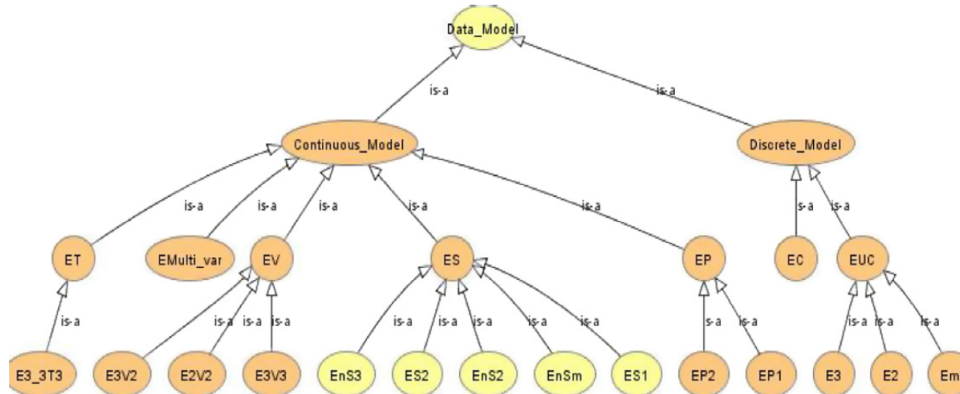
GraphXML is another example for a visualization model provided in XML. GraphXML is a graph description language that serves as an interchange format between graph drawing and visualization applications [HM01]. The simple form of the language describes the basic components of a graph, its nodes and edges respectively. Additional data can be added to the graph description to represent e.g., domain-dependent data or the different types of applications. Moreover, the language supports the description of graph hierarchies, storing the history of the user's

```
<graph id="example">
  <style>
    <line tag="node" linestyle="solid" colour="blue"/>
    <fill tag="node" fillstyle="solid" colour="white"/>
    <line tag="edge" linestyle="solid" colour="blue"/>
    <fill tag="edge" fillstyle="node"/>
  </style>
  <node name="M">
    <position x="30" y="10"/>
    <size width="20" height="20"/>
  </node>
  <!--Similarly nodes A and S-->
  <edge source="M" target="A"/>
  <edge source="M" target="S"/>
</graph>
<edit action="replace" xlink:href="#example">
  <!--Specify the same graph as above with updated positions-->
  <node name="F">
    <position x="50" y="90"/>
    <size width="20" height="20"/>
  </node>
  <edge source="A" target="F"/>
</edit>
```

**Figure 2.9:** GraphXML describing basic components of a graph, its nodes and edges respectively. From [KKMN10], Figure 5.

actions in some form of journaling, storing the geometry (position, size of the node, edge geometry etc.) and the description of the visual properties (line width, color of the components, icons replacing nodes etc.). The Figure 2.9 shows an example for the GraphXML.

Although these models are fairly comprehensive with respect to visual encoding, specifying all basic elements of visualizations (cf., titles, data points, axes, trend lines) and how they must be encoded to be readable in other applications, they fall short in showing the relations between composing visual objects (eg., visual channels) and in providing high-level context information (semantics). Thus, Shu et al. [SAR08] and Duke et al. [DBDH05] suggest to use an ontology language to model the knowledge about visualizations. An ontology represents not only the visual objects involved in the visual mapping but also their relation to each other. It provides more semantic about the composing visual objects that significantly facilitates searching and browsing the visual artefacts and discovering the domain knowledge [SAR08]. The following section discusses the most relevant visualization ontologies.

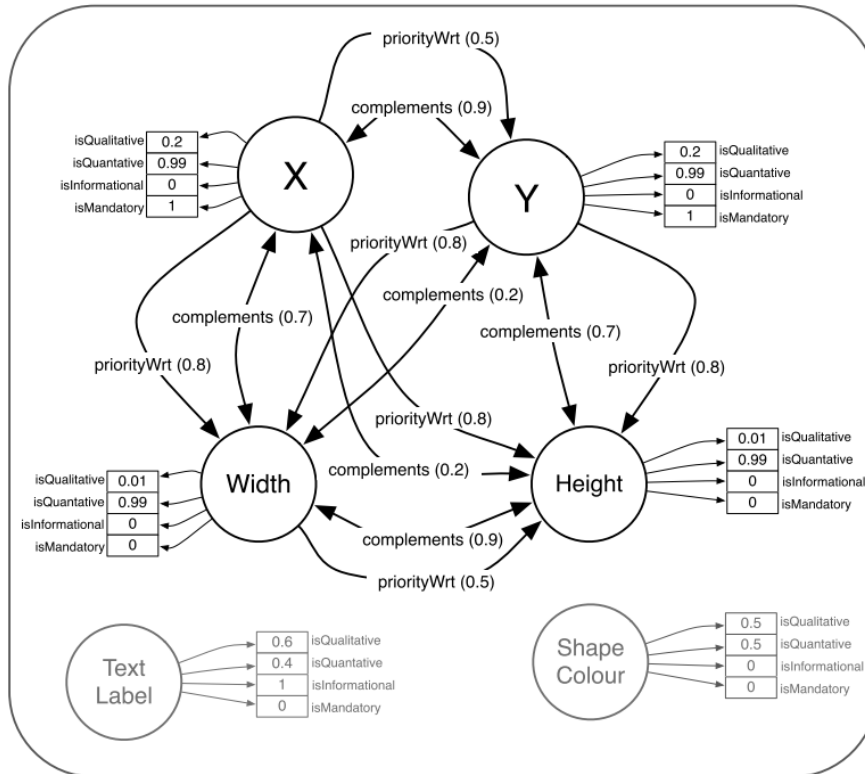


**Figure 2.10:** A fragment of the Visualization Ontology describing the data model. From [SAR08], Figure 2.

### 2.2.1.2 Visualization Ontologies

Shu et al. [SAR08] introduce a Visualization Ontology (VO) which comprises four abstract classes representing the main concept of the visualization domain: the data model, visualization techniques, data representation and set of basic concepts (primitive set) [SAR08]. Moreover, it highlights the connection between them. The data model describes the data and a variety of techniques and algorithms for visualizing the data (e.g., isosurface for a 3D regular sample set or a 2D graph for a series of 2D coordinates). The data representation concept comprises multimodal attributes (visual, haptic, sonic, olfactory, physical attributes etc.) whereby the primitive set contains the building blocks of above mentioned classes. Figure 2.10 shows a fragment of the ontology. In a nutshell, the data model in Figure 2.10 is categorized into discrete- and continuous model. The continuous model is further classified according to the type of each variable: scalar, vector, tensor, point or multi-variate which belong to the class DataSetType, a sub-class of the primitive sets. Finally, each variable is broken down regarding to its dimensionality. The discrete model however is classified into either a connected or unconnected model where the unconnected model is further broken down according to the dimensionality of the data. The aim of the ontology is (i) to share process models between the developers and the users and (ii) to collaborate and interact between distributed sites that offer visualizations services [SAR08].

Yet, this ontology describes the visualizations in a very abstract manner and lacks simple visual encoding facts and guidelines. As a result, it does not sufficiently represent existing knowledge about information visualization. This and the fact that this ontology only focuses on the visualization process and not on the required visual objects needed for the visual mapping, makes it difficult for it to be entirely reused by other visualization tools. Moreover, the ontology comprises just a few simple



**Figure 2.11:** The Visual Representation Ontology illustrating a 2D Graph for music charts. From [GSGC08b], Figure 4.

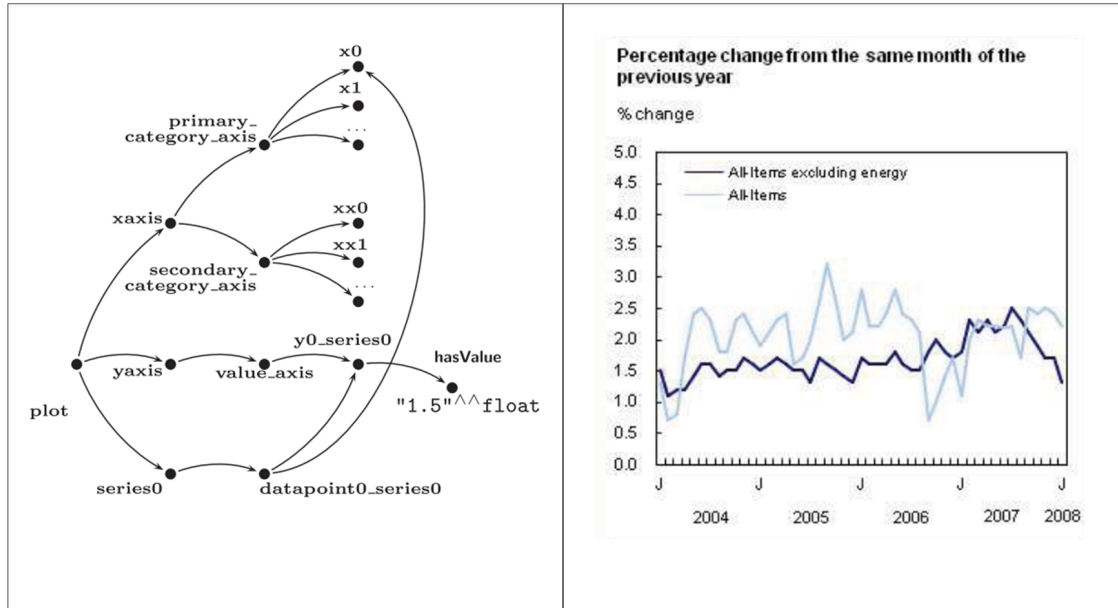
visualizations and is not accessible to the public.

Leida et. al [LXTM11] present a visualization ontology which provides semantic description of bar chart, pie chart and line chart with attributes that define specific parameters for the used visualization library (Google Visualization API <sup>5</sup>, Yahoo UI Charts <sup>6</sup> or Exhibit <sup>7</sup>). Yet, this defines the main element of the ontology: the visualization code associated with each visualization method. What this means is, the ontology is not created to serve as a unified specification instead it focuses on the visualization process and bound to the specific terminologies and taxonomies of certain libraries. Moreover, this ontology is not accessible to the public too.

Gibson et. al. [GSGC08b] propose an approach for automatic generation of visualizations from domain-specific data available on the web using ontology mapping and probabilistic reasoning techniques. To achieve that, the data from a specific subject web page (music charts) is first mapped to a Domain Ontology which is then mapped to one or more Visual Representation Ontologies (VRO), each of which

<sup>5</sup><https://developers.google.com/chart/>  
<sup>6</sup><https://yuilibrary.com/yui/docs/charts/>  
<sup>7</sup><http://simile-widgets.org/exhibit/>





**Figure 2.12:** Diagrammatic view of SGO (see left side) describing a line graph showing the yearly percentage change in e.g., manufacturing sales (see right side). From [DFVR10], Figure 3 & 4.

captures the semantics of a visualization. The ontology is defined by specifying the visual artefacts (e.g., *x - coordinate*, *y - coordinate*, *color* etc.) of the visualizations and their relationship to each other (see Figure 2.11). It is so far used to describe TreeMaps, Parallel Coordinates and Graph Networks and this for a specific use case: visually presenting of music charts. Although this ontology captures the basic properties of 2D visualizations, it is restricted to a single domain and a handful of visualizations. There is no general specification provided for describing different kind of visualizations. As a result, the ontology cannot be applied in individual domains, especially if the domain requires additional visualizations techniques than TreeMaps, Parallel Coordinates and Graph Networks.

The Statistical Graph Ontology (SGO) by Dumontier et al. [DFVR10] describes the structure of the statistical graphs (=visualizations, charts) semantically in order to make these graphs available as highly structured representations that can be queried, exchanged, integrated and whose structure can be extended using Semantic Web technologies. In general, the ontology models the components and relations that essentially constitutes a statistical graph with the focus on n-variable line graphs. Concretely, the modeling comprises the definition of the title, the axes of the graph with their instances, the categorization of the axes (*category\_axes*, *value\_axes*) and the datatype of the *value\_axes* which can be integer, float, real or double. The Figure 2.12 illustrates the diagrammatic view of the ontology (see left side of the figure) defined for a line graph showing the yearly percentage change in

**Listing 2.4:** Description of the Bar chart.<sup>8</sup>

```
viso-graphic:Bar_Chart
  rdf:type owl:Class ;
  rdfs:label "Bar_Chart" ;
  rdfs:subClassOf viso-graphic:Static_Graphic_Representation .
```

e.g., manufacturing sales (see right side of the figure). While SGO provides a sophisticated ground for describing statistical graphs, it does not represent all required concepts. For instance, the definition of the visual channels color and size are missing in the ontology although required to define statistical visualizations such as bar chart, scatter plot or bubble chart. Furthermore, the ontology handles the datatype definition too general: it distinguish only between categorical and numerical data and does not support specialized datatype definitions such as location required for a map visualization or date required for a time visualization. Finally, the ontology is restricted on line graphs and does not provide a general specification to describe other kind of graphs than line graphs including multidimensional graphs such as parallel coordinates, scatter plot matrix or heat map.

Finally, Voigt et al. [VFM13] provide a concept for a visualization ontology which annotates the data and visualization components, contains factual knowledge of the visualization domain (i.e., visual encoding rules, perceptual guidelines) and serves as semantic framework for storing contextual information. Yet, the ontology defines each involved concept (e.g., visual channels, charts etc.) individually and ignores the relation the concepts might have to each other. To clarify this, we consider the Listing 2.4 which illustrates the description of the bar chart. Looking at the description, we can observe that there is no relation defined between the chart and its visual channels (*x-axis*, *y-axis* and *color*). To be appropriate, each chart has its unique requirements and features. Thus, it is not enough to provide a general description of a visual channel but also which role it takes in a visualization. For instance, *color* is a optional channel for a bar chart whereby *x-axis* and *y-axis* are mandatory. Moreover, line chart uses *x-axis* to plot temporal or quantitative data whereby bar chart uses it to plot categorical data. Summarized, this ontology is not the best choice to describe visualizations in a general way including the related visual object and their properties.

The remaining ontologies are either tailored to specific software visualization systems [RKR06] or visualization algorithms [TM04] both not contributing to a domain independent visualization ontologies.

---

<sup>8</sup><https://github.com/viso-ontology/viso-ontology/blob/master/src/main/resources/modules/graphic/viso-graphic.ttl>

### 2.2.1.3 Summary and Difference to the State-of-the-art on the Semantics of Encoding Data Visually

The literature review showed evidence that there is a lack in providing a general model for visualizations that:

- comprises the most important expert knowledge about information visualization required to create appropriate visualizations
- describes the relation between the visualizations and their components (e.g., visual channels)
- provides semantics to the visualizations and their components
- comprises multidimensional visualizations
- is extensible with new visualizations
- is domain- and technology independent
- is machine- and human readable

In Chapter 3 we provide the Visual Analytics (VA) Vocabulary representing visualizations (including multidimensional visualizations) in a common persistence model that can be reused by various technologies. The VA Vocabulary is an explicit conceptualization that describes visualizations in pragmatic simple facts considering visual encoding rules and perceptual guidelines that will aid the sensible mapping from data to the appropriate visualizations. Further, the vocabulary defines the specification of our visualization ontology that provides semantics to the visualizations. Together with the mapping algorithm, this ontology builds the rule-based recommender which automatically creates the list of appropriate visualizations for a given dataset.

## 2.2.2 Visualization Recommendation

A large number of systems has been defined to automatically recommend and create appropriate visualizations. Existing systems can be categorized regarding to the factors they consider: data characteristic, intended task, semantic and domain knowledge represented in the data and the user preferences [VHS<sup>+</sup>15].

Data characteristic-based visual recommender systems focus on effectively expressing the desired information from user's data. This includes, choosing the most effective method for (i) encoding the information (Expressiveness criteria) and (ii) their visual representation (Effectiveness criteria) [Mac86]. Yet, there is a set of facts required to address (i) and (ii) which comprises visual encoding rules and perceptual guidelines. As shown in Section 2.2.1, such rules and guidelines can be expressed in a language that describes “the syntactic and semantic properties” [Mac86] of the

visualizations and can be reused to define which visualizations are more accurate for a given dataset. In general, the formalization of visual encoding rules and principles not only improves our understanding of the visualization process, it also contributes to the formulation of generative methods for automated visualization, rule-based systems respectively [MVT16].

A task-based visual recommender system aims to identify user's goal behind visualizing the data [VHS<sup>+</sup>15]. User's goal can be categorized into two groups: exploratory- or confirmatory. The former describes an analysis process where the user aims to summarize the main features of the data and use them to define a preliminary hypotheses. The later, however, describes the process where the user tests her found or assumed hypotheses from the previous analysis. Basically, the task-based visual recommender analyzes user's behavior to infer her intended task. However, some systems directly ask the user to share her goal with the system. This information is then used by the recommender to reduce the list of visual recommenders on those which the system assumes would better address user's tasks.

A semantic-based recommender uses the semantic information provided with the data to define recommendations. Such information might describe the content of the data, the characteristics of the data attributes, the relation between the data attributes and between the current and other datasets. The domain- based recommender however uses the unique domain knowledge to recommend visualizations which show e.g. the trends, patterns and outliers within the data [VHS<sup>+</sup>15, KO17].

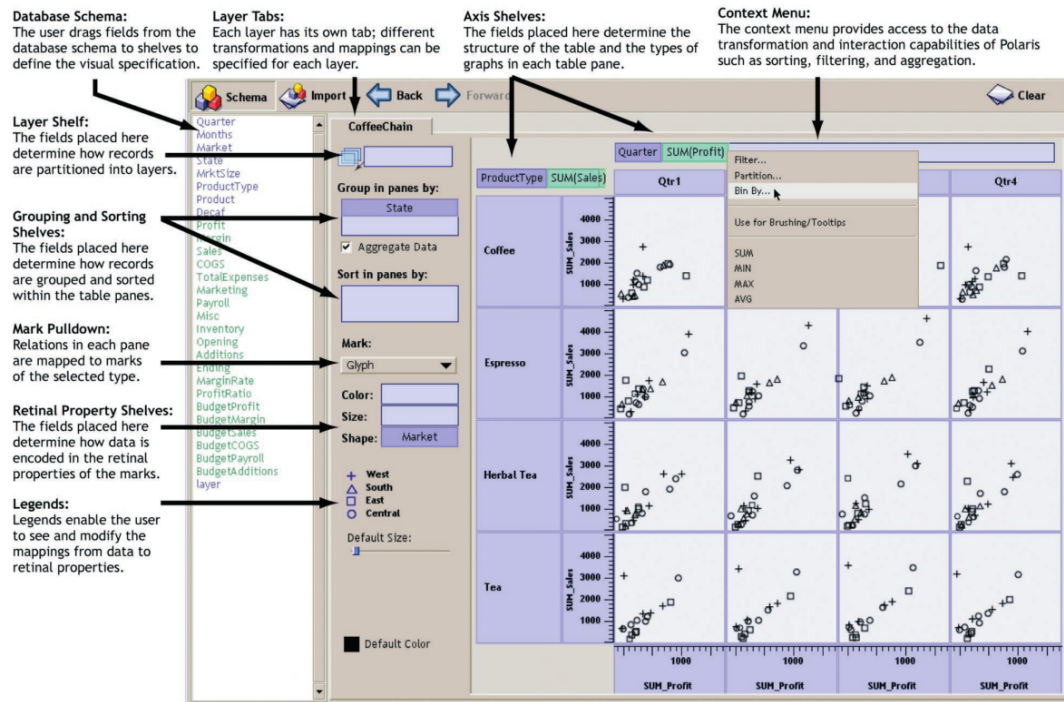
Finally, visual recommenders based on the user preferences define recommendations considering user's visual taste, interest or needs.

The following sections investigates existing visual recommenders regarding the above mentioned factors. The literature review is restricted to the years between 2002 and 2016. However, we do not claim completeness for all publications available in this area. Note, these sections are mainly based on the publication P7 [MVT16].

### 2.2.2.1 Rule-based Approaches

Following Mackinlay [Mac86], the initial referent for the automated generation of visualizations is Polaris, the backbone engine in the early version of *Tableau* [STH02]. The system automatically suggests visualizations for tables in relational databases and coordinates the interaction between them. But, the mapping of data onto visual components of a visualization is not performed automatically, instead it has to be formulated by the user (see Figure 2.13).

Cammarano et al. [CDC<sup>+</sup>07] provides a system which follows a bottom-up approach to automate the visualization process. To visualize the data, the user first formulates a query (free-form text common in search engines), for which she obtains a set of relevant objects (equivalent to RDF *object*, see Section 2.1.2). Next step includes the selection of one of the supported visualizations: map, timeline or scatter-plot. Each visualization is specified by a set of triples  $\{(T_1, N_1, E_1), \dots (T_k, N_k, E_k)\}$ , where  $T_i$  represents the type of attribute,  $N_i$  the name of the attribute and finally  $E_i$



**Figure 2.13:** The Polaris user interface. To create a visualization, the user first selects the desired fields from the database schema onto shelves. The system then performs the required visual operations to visually display the values of the selected fields. From [STH02], Figure 1.

the visual encoding for the attribute. Once selected a certain visualization, the system then searches for the attributes in the data space that best fit the requirements of the chosen visualization. To clarify this, consider the following example. Assume, the user wants to see the residences (states) of the U.S senators on a map. The corresponding fields to create such a map are the senator's picture and the names and the geographical coordinates of their associated states. The visualization specification given to the search algorithm thus is:

$$\{(\text{decimal}, \text{state latitude}, y),$$

$$(\text{decimal}, \text{state longitude}, x),$$

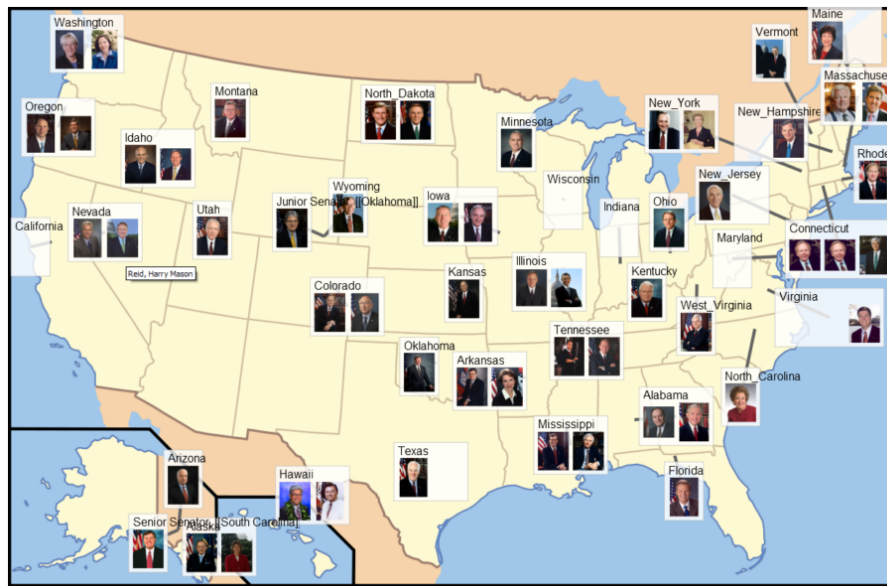
$$(\text{string}, \text{name}, \text{tooltip}),$$

$$(\text{Img}, \text{image}, \text{icon}),$$

$$(\text{string}, \text{state}, \text{text})\}$$

Figure 2.14 shows the produced visualization for this query.

Mackinlay et al. propose an influential, albeit conceptually different approach, in the ShowMe [MHS07] system. It integrates a set of the user interface commands

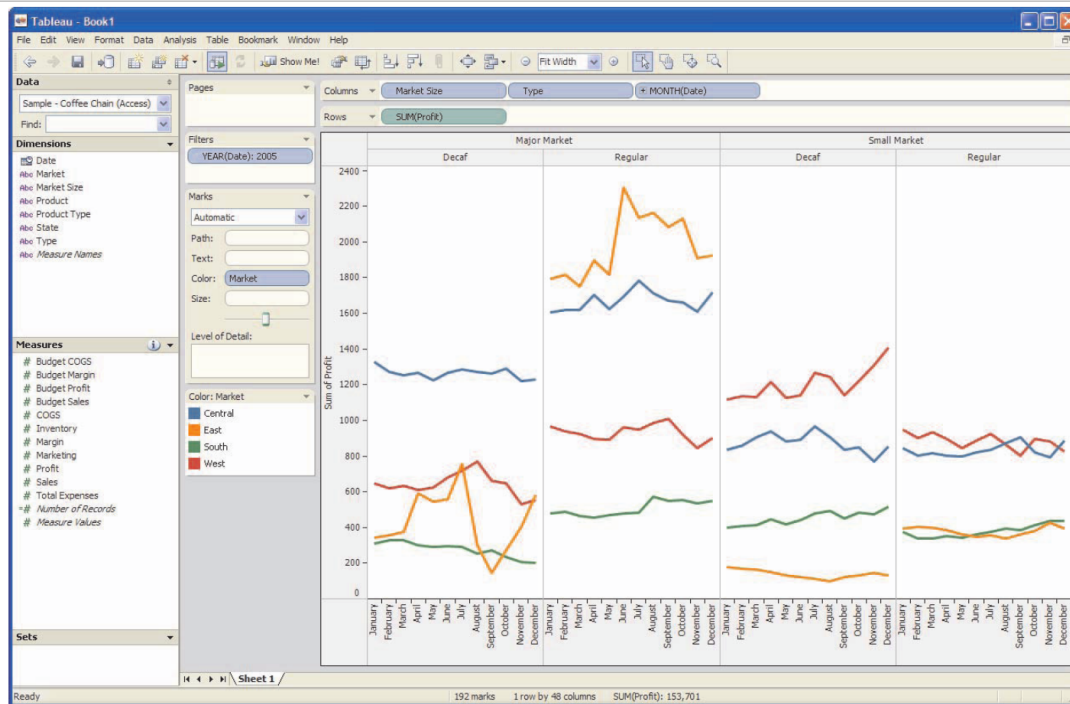


**Figure 2.14:** A map showing the residences of the U.S senators. From [CDC<sup>+</sup>07], Figure 2.

and functions aiming at automatically generating visualizations for the commercial tool *Tableau*.<sup>9</sup> ShowMe attempts to help the user by searching for graphical presentations that may address her task. For creating a visualization, the user first drags the fields of interest from so called data window on the so called shelves (row, column), which organizes the fields regarding to their type: in dimensions or measures respectively. After this, the system selects the most appropriate visualization for the selected fields and presents the data using best practices of graphic design (see Figure 2.15). The visualizations are selected based on the data properties, such as datatype (text, date, time, numeric, boolean), data role (measure or dimension) and data interpretation (discrete or continuous). Similar to the approach defined by Cammarano et al. [CDC<sup>+</sup>07], this system falls short of automating the whole process and leave decisions that require certain expertise to the user.

Vispedia is an interactive Web-based visualization system [CWT<sup>+</sup>08] to visualize heterogeneous datasets. The visualization process of Vispedia is based on the integration of the selected data (tabular data) into an iterative and interactive data exploration and analysis process enabling non-experts to more effectively visualize the semi-structured data available [MHS<sup>+</sup>13]. Figure 2.16 illustrates the workflow of Vispedia. In a nutshell, on a certain Wikipedia page the user first selects a table using the Vispedia bookmarklet. As next, she picks one of the three visualizations for which she has to formulate an initial search query that comprises the appropriate fields for the selected visualization. Once the system detects the data matching the

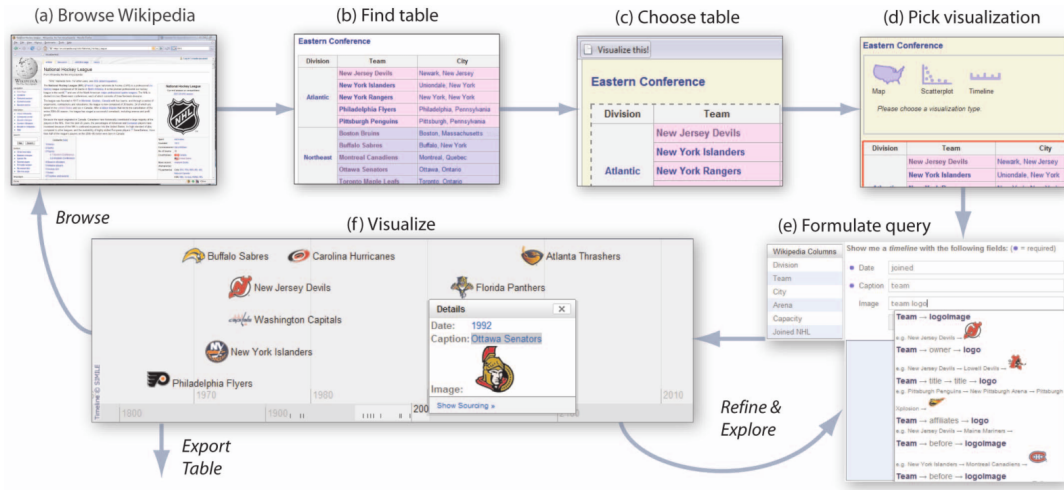
<sup>9</sup><http://www.tableausoftware.com/>



**Figure 2.15:** The ShowMe user interface. For creating a visualization, the user first drags the fields of interest from so called data window on the so called shelves (row, column), which organizes the fields regarding to their type: in dimensions or measures respectively. After this, the system selects the most appropriate visualization for the selected fields and presents the data using best practices of graphic design. From [MHS07], Figure 1.

query, it finally creates the candidate visualization. Although effective, the Vispedia is limited to only Wikipedia data and requires manually formulated mapping definitions [STV<sup>+</sup>14].

Many Eyes [VWV<sup>+</sup>07] is a public web site to upload data, create interactive visualizations and share them with the other users. The later should foster the collaboration among the users in terms of commenting and annotating uploaded visualizations which is the main focus of the tool. Although powerful, ManyEyes falls short in automating the visualization process. To set up a visual mapping, the user combines the visualization with the attributes of the data manually. To achieve this, a schema specifies the data needs of a visualization using a set of named, typed slots (see Figure 2.17). In this schema, “T” denotes textual data, “N” numeric data, “T+” multiple textual data whereby “N+” denotes multiple numerical data. For instance, using this schema, the data needs of a treemap or a scatterplot can be expressed as follows:  $\{hierarchy : t+, size : N, color : N\}$  and  $\{Xaxis : N, Yaxis : N, label : T, [Dotsize : N]\}$ . Yet, the challenge for the user is to map the correct data attribute to a slot and finally create a meaningful visualization.



**Figure 2.16:** The Vispedia Workflow. On a certain Wikipedia page the user first selects a table using the Vispedia bookmarklet. As next, she picks one of the three visualizations for which she has to formulate an initial search query that comprises the appropriate fields for the selected visualization. Once the system detects the data matching the query, it finally creates the candidate visualization. From [CWT<sup>+</sup>08], Figure 2.

SEEDB [VMPP14] is a system which automatically identifies and recommends the most interesting visualizations for the current query results using methods based on probability distribution, distance matrices and deviation. Concretely, each visualization generated automatically shows a certain aspect of the current dataset. Yet, when a user selects a visualization the system recommends alternative ones that show an aspect of the dataset that deviates from them shown on the selected visualization. To clarify this, consider the example illustrated in Figure 2.18. In this example, the user was interested in the “total sales by store”. The Figure 2.18a shows the visualization the user currently explores. This visualization illustrates the “Laserwave sales” by store. Yet, the system recommends the user an alternative visualization (see Figure 2.18b) that shows the “overall sales of the products” since they follow an opposite trend compared to the “Laserwave sales”. In contrast, the same visualization (see Figure 2.18c) would not be recommended when the sales of all products would follow a similar trend like the “Laserwaves sale” from Figure 2.18a. This system has two limitations. First, it only supports a two-column table and two basic charts, bar chart and line chart respectively. Second, when having huge datasets a real time exploration of the entire dataset and all possible visualizations with regard to their deviations from each other might cause huge scalability issues.

Voyager [WMA<sup>+</sup>16] is a recommender engine which generates visual recommendations based on the computed data characteristic and the user-specified attributes of interest. It has been developed to assist the users when visualizing their data. Once the user selects attributes to visualize, the system automatically recommends



Technique	Data schema
Bubblechart	
Histogram	
Pie Chart	{Labels / item names : T, Values : N}
Maps	
Tag Cloud	
Bar chart	
Line graph	{Axis labels : T, Values : N+}
Stack graph	
Network diagram	{From : T, To : T}
Scatterplot	{Xaxis : N, Yaxis : N, Label: T, [Dotsize : N] }
Stack graph/ categories	{Hierarchy : T+, Values : N+}
Treemaps	{Hierarchy : T+, Size : N , Color : N}
Tag Cloud	{U}

**Figure 2.17:** Many Eyes Schema for visual mapping. From [VWV<sup>+</sup>07], Table 1.

additional attributes that the user might otherwise overlook. Based on the selected and recommended attributes, the system then defines a set of visual recommendations and has the user select among them. The following example describes this process. Assume user's data is about *cars* and she is interested in knowing the *horsepower* of each car. When the user now selects the attribute *horsepower*, the recommended visualizations would show some suggested attributes in addition to horsepower (see Figure 2.19). The main characteristic of Voyager is, it prioritizes data variation (different variables and transformations) over design variation (different encodings of the same data).

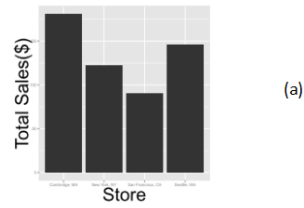
### 2.2.2.2 Behavior-based Approaches

Nazemi et al [NRB<sup>+</sup>13] propose a system which tracks user's behavior (interactions on visualizations) in order to adapt a set of applicable visualizations on user's intention behind visualizing the data. Regarding to Nazemi et al., there are four possible information seeking goals where the users have to navigate through their data to perform a deep analysis. These goals include searching, locating, analyzing and exploring. Yet, when performing one of these tasks, the proposed approach suggests to adapt the visual presentations on the different level of user's goal requirements and not only on the data characteristics. The user behavior is investigated applying a canonical user model. However, the system is also capable of responding to the individual user. To do so, Nazemi et. al. utilize a personalized approach to recommend visualization, but only target the content from digital libraries (i.e., bibliographical notes and publications).

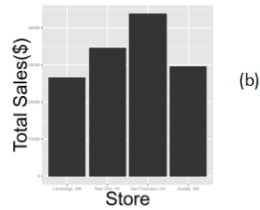
Ahn et al. work on VIBE [AB09], an adaptive visualization tool, that attempts to provide the user-adapted visual representation of their search results. The user

Table 1: Data: Total Sales by Store for Laserwave

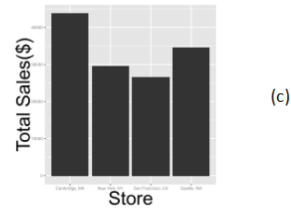
Store	Total Sales (\$)
Cambridge, MA	180.55
Seattle, WA	145.50
New York, NY	122.00
San Francisco, CA	90.13



Visualization: Total Sales by Store for Laserwave



Scenario A: Total Sales by Store



Scenario B: Total Sales by Store

**Figure 2.18:** SEEDB: (a) Shows user’s current visualization. (b) Shows the visualization that has been recommended by the system since it shows data that deviate from them shown in (a). (c) Shows a visualization that would not be recommended since it follows a similar trend like (a). From [VMPP14], Figure 1, 2 and 3.

context is a collection of the user actions accumulated over time, such as the issued search queries, selected documents from the search results and traversed links. The collection captures user’s interests beyond the query and in turn defines a user model which is applied to visually highlight the relevance of a particular result set (=documents). The Figure 2.20 illustrates the document placements based on their similarity to the queried term and the user model of the current user. The yellow circles illustrate the queried term whereby the blue circles denotes the extracted named-entities and represent the user model. The resulted documents are represented as squares whereby their position indicate their similarity to the queried term or the user model [Naz16]. Although VIBE is a promising system that supports the user during their search process, it only uses a single visualization and does not provide recommendation capabilities.

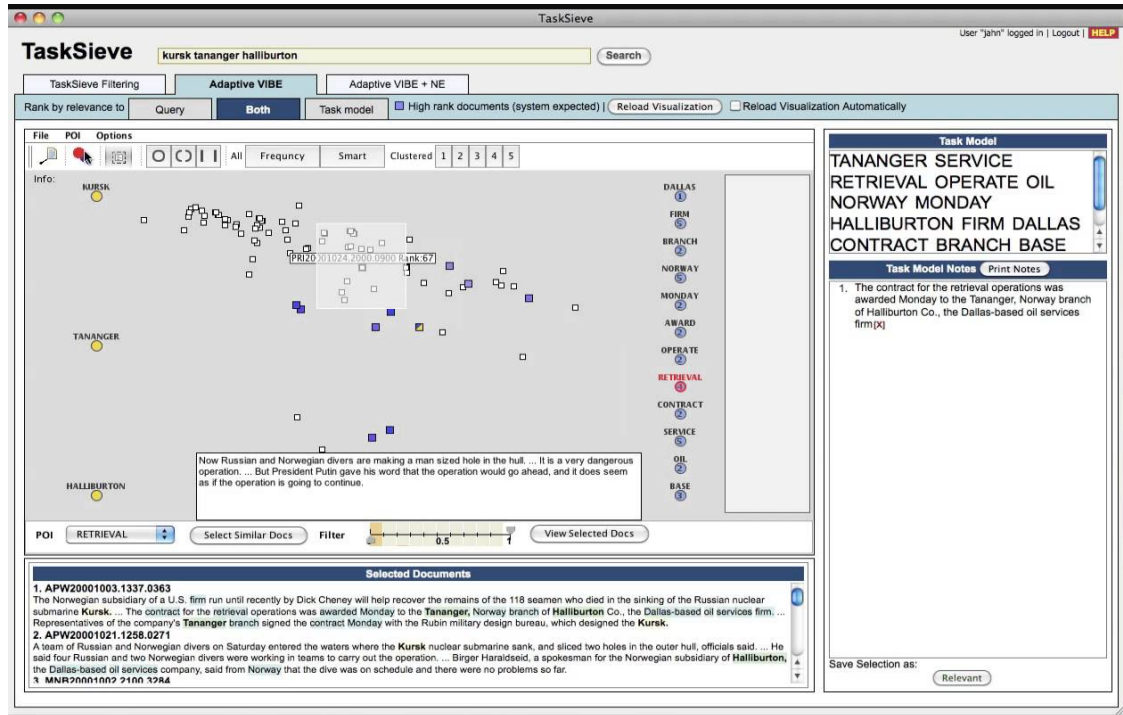
The system by Gotz et. al [GW09] also attempts to infer user’s intended task considering her behavior. This information is used by the system to recommend suitable visualizations that might support the user in performing her current task. Concretely, the system holds a library of behavior patterns defined by experts through observations of real-world visual analytic activity [GW09]. To create the visual recommendations, the system tries to match user’s current behavior to one of those predefined patterns. Once a pattern is detected, the system uses it to infer user’s



**Figure 2.19:** The Voyager user interface: The schema panel (left) lists the user-selected attributes. The main gallery (right) presents recommended visualizations of different attribute subsets and transformations. From [WMA<sup>+</sup>16], Figure 1.

intended task and recommend alternative visualizations that might support the task more directly than user’s current visualization [GW09]. Figure 2.21 illustrates the user interface of the behavior-driven visualization recommender. The approach by Gotz et al. is interesting in that it supports the users in complex visual analytical tasks. However, the patterns they use are defined by experts a priori, rather than based on the results of experimental observations reflecting the individual user characteristics [SCC13]. Moreover, the system is limited on one single visualization and the repeated patterns change only the layout of the single visualization.

Finally, VizDeck [KHPA12] generates visual recommendations based on log data collected by the system and the statistical properties of the data. The log data comprises users’ interactions (promotions, discards, filter events, highlight events) on the recommended visualizations and are used by the system to statistically learn which visualizations are more appropriate for the current dataset in terms of addressing user’s visual needs. Each time a user interacts with a visualization, the system extracts the statistical features of the data such as number on distinct values, entropy, coefficient of variation, kurtosis and periodicity. This information is then applied to generate more appropriate recommendations when similar data is provided in the future.

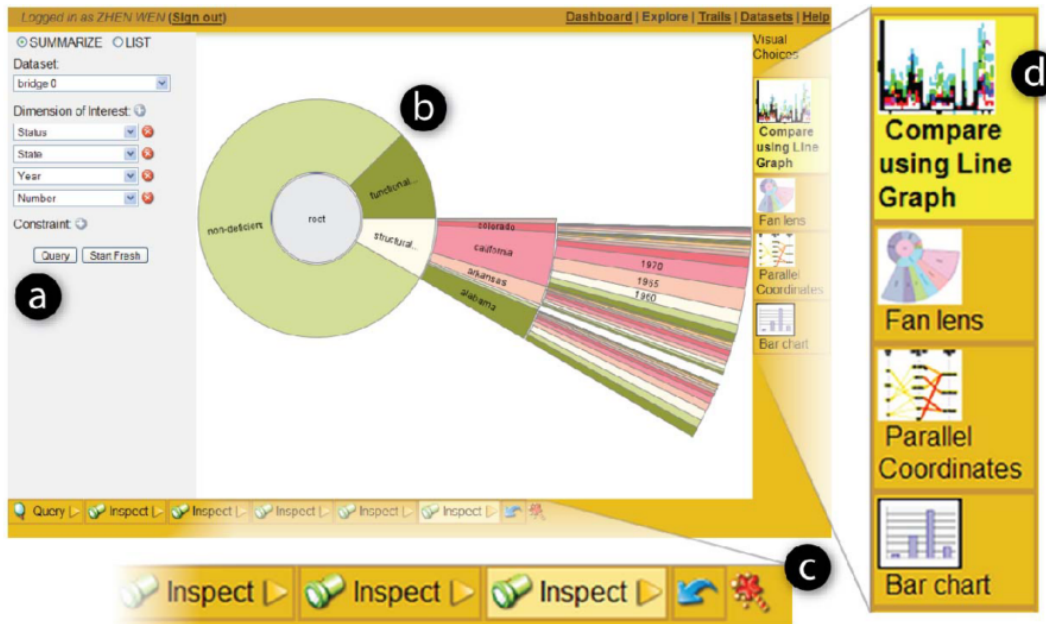


**Figure 2.20:** The adaptive VIBE showing the document placements based on their similarity to the queried term and the user model of the current user. The yellow circles illustrate the queried term whereby the blue circles denotes the extracted named-entities and represent the user model. The squares represent the resulted documents whereby their position indicate their similarity to the queried term or user model. From [AB09], Figure 2.

### 2.2.2.3 Semantic-based Approaches

Gilson et al. [GSGC08a] provides a system which automatically recommends visualizations for domain-specific data available on the web using ontology mapping and probabilistic reasoning techniques. To achieve this, the data from a specific subject web page (music charts) is mapped to a Domain Ontology and the visualizations are structured in a Visual Representation Ontology (VRO) (see Section 2.2.1.2). To perform the visual mapping, a third ontology, the *Semantic bridging ontology*, is used that bridges the information from two ontologies and finally creates the appropriate visualizations. However, the resulting visualizations are tailored to a single platform and do not offer a dynamic solution.

Voigt et al. propose a system [VFM13] which uses a knowledge base of numerous ontologies to recommend visualizations. It is essentially a rule-based system that preselects visualizations based on the device, data properties and task involved. Subsequently, the system ranks visualizations following the rules concerning visualization facts, domain assignments and the user context that has been presented in



**Figure 2.21:** Behavior-driven visualization recommender. To use the recommender, the user first defines a query (a). Next, the system generates a visualization the user can interact with to analyze her data (b). Based on user’s current behavior, the system defines visual recommendations, notify (c) the user via “magic wand” icon in the history panel and (d) a flashing segment on the recommendation sidebar. The user can either accept the recommendation by clicking on the icon or deny it. From [GW09], Figure 1.

form of ontologies. Although promising, this approach is mostly theoretical. While the user context information and visualization literacy are outlined as required in this approach, they mostly proposed a concept where the implementation and data collection parts are missing.

#### 2.2.2.4 User Preferences Oriented Approaches

Steichen et. al [SCC13] present research that aims to support the adaptive visualization systems to address each individual user’s needs and interest in real time. Basically, the research focuses on predicting user’s visual tasks of varying type and complexity to suggest an alternative visualization which reflects more accurately what the user needs and prefers than the current visualization. To achieve this, the system uses the information on eye gaze patterns while a user interacts with a given visualization. These patterns have been defined based on the observations made in a previous lab experiment where the participants had to perform nine tasks from varying type and complexity using a certain visualization. Concluded, this system generates visualization suggestions using information that is accumulated from the

eye movement patterns of a user while viewing a visualization. This however makes the proposed approach highly interface-dependent as the use of an eye-tracker is essential in order to collect the gaze information.

### 2.2.2.5 Summary and Difference to the State-of-the-art

The relevant work reported within the scope of visualization recommender system proves, the requirements such systems should address changed and expanded within the last years [KO17]. Starting with only considering the data characteristics and domain specific knowledge, the recent visual recommenders focus on a user-centric design in order to reflect user's task and needs more accurately. To achieve this, the systems either require the user to specify her goal manually or observe and analyze user's behavior to infer her intention behind visualizing the data. This information is used to create a user model. The user model, however, is only one source for defining visual recommendations [MTD<sup>+</sup>07]. The initial visualizations are primarily guided by existing knowledge about visual encoding rules and perceptual guidelines using, for instance, a rule-based system. This should ensure that the most important attributes are encoded with the most effective visual channels.

Basically, a rule-based systems use an organized taxonomy (or ontology) describing the visual encoding rules and guidelines to map the data to the corresponding visual components. As a general issue, the existing taxonomies (or ontologies) are either domain or technology dependent or include only a small subset of possible visualizations. Beside that, existing rule-based systems build upon taxonomies or ontologies are mostly difficult to use [GW09], requiring manual specification of visual information which involves the following: selecting variables of interest, selecting transformation and designing encodings [MVT16]. All together is a tedious task that interrupts the exploration flow.

In this thesis, we propose a visual recommender that improves the existing approaches along two main directions:

- providing a common, persistence model for visualizations that provides deep semantics and can be reused by various technologies and domains
- automatically defining a list of appropriate visualizations using the above-mentioned model.

Using a common description of visualizations, our visual recommender “makes an intelligent use of human visual abilities and output media” [Mac86] that support the user in her exploratory analysis more precisely. However, the main feature of our recommender is, it personalizes the list of appropriate visualizations considering user's visual needs and interests before presenting it to her. To do so, we apply conventional recommendation approaches collaborative filtering and content-based filtering (see Section 2.1.3). As a result, the recommended visualizations do not only express the characteristics of user's data but also her preferences.

In contrast to the existing work in adaptive visualizations that work bottom-up, analyzing the user actions to determine her behavior and thereby predict the desired configuration of the visual display, our system is build on explicitly expressed preferences provided in form of ratings or tags. This is mainly because, considering only user’s past interactions, usual in behavior based approaches, often causes over-generalization and drives predictions which “do not correspond with the choices, the user would make in a particular situation” [MTD<sup>+</sup>07]. In contrast, we propose a system which is based on user’s current feedback that are implicitly provided in form of ratings describing user’s visual taste and/or tags describing user’s topic of interest.

Our approach for recommending visualizations has already served as an inspiration for some recent research in Information Visualization. For instance, Kanchana et al [KMM<sup>+</sup>16] introduce a paradigm for data visualization using machine learning based context identification. The proposed framework recommends the most appropriate visualizations by considering the context of each dimension of the given dataset. Although promising, this approach ignores the preferences of the individual users. Another approach recommends visualizations based on a multidimensional characterization (nominal, ordinal, quantitative, etc.) of the individual variables but, similar to the previous approach, does not consider the individual user preferences [MMVM0]. Finally, Thiele [TVS<sup>+</sup>17] et al. emphasize the benefits of interactive visualizations tools in analyzing large bodies of text by presenting VizRec as an promising example therefor.

### 2.2.3 Crowd Sourcing Visualization Studies

The data describing the preferences of the users were collected in a crowd-sourced visualization study. We used this data to generate personalized recommendations using our visual recommender (see Chapter 5). A concern with crowd-sourced studies is the lack of control over many experimental conditions, which may impact ecological validity. Nevertheless, perception studies in crowd-sourced platforms are viable, as evidenced by a growing number of successful studies in visualization and related fields [KCS08, HB10b, BVB<sup>+</sup>13, LFK<sup>+</sup>13]. Borkin et al. [BVB<sup>+</sup>13] investigated memorability of visualizations. Considering visualizations much like a static picture, they performed a crowd-sourced study to determine which types of visualizations are better recalled. Investigating perceptual aspects of visualizations, Heer et al. replicated the influential experiments of Cleveland and McGill in the format of a crowd-sourced study [HB10b]. Lin et al. performed a crowd-sourced experiment to determine semantically resonant colors, that is colors that people associate with entities or effects. They derive guidelines for using color in visualization [LFK<sup>+</sup>13]. Carefully designed tasks are mandatory to elicit valid data from crowd platforms. Kittur et al. discuss several design considerations for developing the tasks in crowd-sourced studies [KCS08]. One design recommendation is to have explicitly verifiable questions as part of a task. They found that asking tags for the content is useful

---

because it requires the users to process the content. Hence, we used this guideline to setup our crowd-sourced study designed to elicit the user preferences related to automatically generated visualizations. This study is described in Section 5.3.

## 2.3 Summary

This chapter provides a frame of the topics addressed in the thesis, derived from visualization modeling, visualization recommendation and elicit the user preferences for personalized visualization. In general, it identifies the building blocks for (i) formulating visualizations and visual components and (ii) automated visualization recommendation that guide the creation of a novel approach described in the following chapters. Thereafter, the state-of-the-art research related to this thesis is presented in order to define the frame of the thesis.

The next chapter will detail the workflow and propose a fully functional infrastructure for the visualization recommender including a modeling vocabulary for visualizations.





# Chapter 3

## Automated Visualization Recommendation

This chapter details the Visual Analytics Vocabulary and the rule-based visual recommender that is, using this vocabulary, able to automatically derive visual recommendations. It starts by introducing the context and the building blocks of the vocabulary. Thereafter, the concept, workflow and the infrastructure of the rule-based recommender is introduced, followed by a scenario demonstrating the use of the recommender for automatically visualizing statistical data extracted from scientific publications.

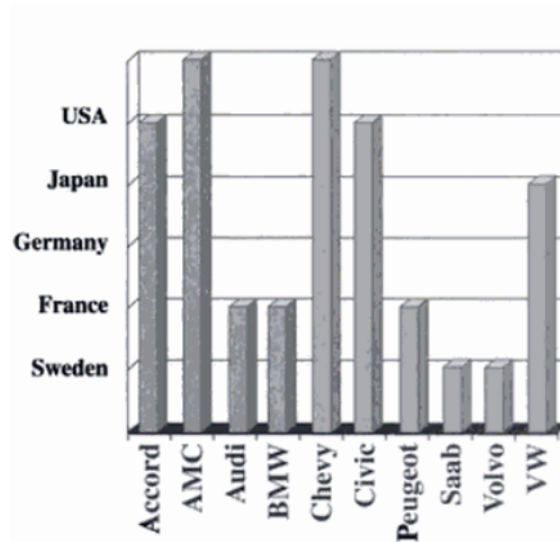
### 3.1 Visual Analytics Vocabulary

This section takes a closer look at the Visual Analytics Vocabulary, analyzing its main concept and building blocks. In particular, the focus is drawn upon the role of the vocabulary in defining visual recommendations.

This section contributes 1) the unified formalization of visualizations using expert knowledge about visual encoding rules and perceptual guidelines, 2) a visualization ontology afforded by the OWL ontology language that describes the semantics of the visualizations and their relationships, 3) visual patterns that originate from the visualization ontology and define, together with the mapping algorithm (see Section 3.2.1.2), the building block of our rule-based system that automatically generate visual recommendations.

#### 3.1.1 Formalization of Visualization Knowledge

“In visualization, the data tables are mapped to visual structures, which augments a spatial substrate with marks and visual channels, to encode information” [CMS99]. There exist multiple ways to perform a mapping but not all defined mappings are perceptually correct. An appropriate mapping is the one which reflects Mackinlay’s



**Figure 3.1:** Incorrect Bar chart encoding two categorical data. From [CMS99], Figure 1.25.

*Expressiveness* and *Effectiveness* principle (see Section 2.1.1.3) so that the information being visualized can be interpreted by the human quickly and correct (see Section 2.1.1.3). However, this is only possible when perceptual characteristics of individual marks and visual channels are considered for visualization process. For instance, the Bar chart illustrated in Figure 3.1 is not expressive since it uses categorical data for both vertical spatial position channel and horizontal spatial position channel and express a nominal relationship which does not exist [CMS99].

The Visual Analytics (VA) Vocabulary is an explicit conceptualization, a model, that describes the visualizations in pragmatic, simple facts that guides the sensible mapping from data to visual representations. The vocabulary has been inspired to a large extent by Bertin’s theory on information visualization and Mackinlay’s formal language to generate graphical presentations (see Section 2.1.1.2). To do so, it unifies the general rules and guidelines in information visualization and provides a basis for creating meaningful visualizations automatically. This includes, the definition of visual mapping rules by a set of relational operators and the visual properties (e.g., visual channels, see Section 2.1.1.2), concepts and characteristics of individual visualizations.

Concretely, the VA Vocabulary consists of two parts:

1. the model of an abstract visualization specifying structural components that any concrete visualization has
2. the model of a concrete visualization refining the abstract visualization model by reification of the visual channels.

**Listing 3.1:** Description of the Bar chart using the VA Vocabulary.

```

va:Chart a rdfs:Class, owl:Class;
  rdfs:label "Chart"@en;
  va:hasChartName va:ChartName;
  va:description "Is_the_graphical_representation_of_numerical_or_qualitative_data."@en;
  va:hasVisualChannel va:VisualChannel;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:VisualChannel a rdfs:Class, owl:Class;
  rdfs:label "Visual_channel"@en;
  rdfs:comment "Represents_a_visual_dimension_of_a_chart."@en;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
  va:hasDataType va:DataType;
  va:hasOccurrence va:Occurrence;
  va:hasPersistence va:Persistence;
.

```

The abstract model comprises (a) *name*, (b) *visual channels* and (c) *description* of the visualizations, the basic visual elements and attributes of the visualizations respectively. However, this information alone might be insufficient in order to avoid visual confusion. For instance, the Figure 3.1 shows a bar chart which encodes two categorical data and thus leads to an incorrect conclusion [CMS99]. One possibility to ensure an optimum encoding is, defining the characteristics of the visual channels. The VA Vocabulary specifies the visual channels by providing the following information:

- *datatype*: set of primitive datatypes that a visual channel supports
- *occurrence*: cardinality, i.e. how many instances are allowed for the visual channel
- *persistence*: whether a visual channel is mandatory for the concrete visualization.

This information should guide the system in choosing the most appropriate visual channel to display a certain data aspect.

The core specification of the Vocabulary is illustrated in Listing 3.1. The first part of 3.1 highlights the general description of visualizations whereas the second part takes over the specification of the visual channels.

In general, using this specification a visual designer is able to describe any scientific visualization type with respect to the visual encoding rules and perceptual guidelines. An exemplary usage of the Vocabulary with its main elements is illustrated in Listing 3.2. In a nutshell, this example demonstrates the description of a bar chart which, regarding to the existing knowledge, has three visual channels, *x*, *y* and *color* each of which holds its own unique properties. Concretely:

- each visual channel supports a specific datatype
- the visual channels  $x$  and  $y$  are mandatory,  $color$  is optional
- each channel has to be instantiated only once to be able to produce valid mapping combinations.

### 3.1.2 Visualization Ontology

The presented knowledge in VA Vocabulary is not directly usable for rendering by the machines. To enable this and further explicitly represent semantics about the visualizations, we applied the defined schema to create a visualization ontology. An excerpt of the ontology is already illustrated in Listing 3.2. As shown in Listing 3.2, the ontology identifies and categorizes each visualization type with respect to its context. Furthermore, it explicitly defines the relation between visualizations and its visual channels. All together important inputs to the search and browsing the visual artefacts.

Currently, the VA Vocabulary is used to describe eleven types of visualizations—Bar chart, Pie chart, Scatter plot, Bubble chart, Stream graph, Line chart, Grouped Bar chart, Map, Scatter plot matrix, Parallel coordinates and Timeline. However, additional visualizations can be integrated in a straightforward way by just following the specification of the vocabulary as shown in Listing 3.1. Moreover, the ontology is domain- and technology independent, provides support for maintainability and is accessible to public.<sup>1</sup>

### 3.1.3 Visual Pattern

Visual patterns result from the fact that, depending on the properties of a visualization, a visual channel can support different datatypes in different combinations. Having described visualizations in terms of visual channels and supported datatypes, visual patterns can be derived, each describing one possible configuration of a visualization [MHT<sup>+</sup>14] (see Figure 3.2). In other words, the patterns specify the types of data that are required for each visualization to be instantiated. For instance, following bar chart description in Listing 3.2, two possible patterns for the bar chart are (1)  $\{x - axis : string, y - axis : number\}$  and (2)  $\{x - axis : date, y - axis : number\}$  representing the fact that the  $x - axis$  can accept both types of data but not at the same time.

In addition, the presence of an optional visual channel results in two additional patterns, concretely (3)  $\{x - axis : string, y - axis : number, color : string\}$ , and (4)  $\{x - axis : date, y - axis : number, color : string\}$ . These patterns will be instantiated only if there is a value that exists for  $color$ , otherwise the system would select patterns (1) and (2).

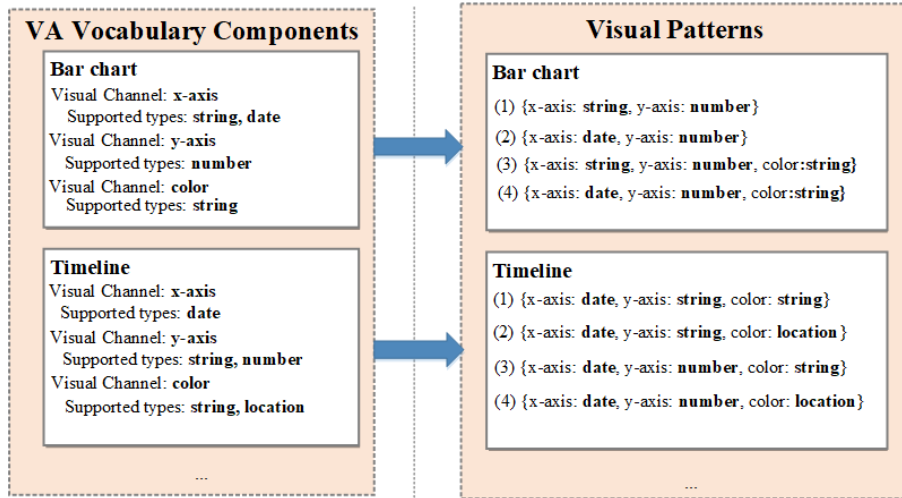
<sup>1</sup><https://code.know-center.tugraz.at/static/ontology/visual-analytics.owl>

**Listing 3.2:** Description of the Bar chart using the VA Vocabulary.

```

<rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:va="http://code-research.eu/ontology/visual-analytics#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <va:BarChart rdf:about="http://code-research.eu/ontology/visual-analyticsBarChart" >
    <rdfs:label>barchart</rdfs:label>
    <va:hasVisualChannels>
      <va:Axis rdf:about="http://code-research.eu/ontology/visual-analyticsBarChartXAxis" >
        <rdfs:label>x-Axis</rdfs:label>
        <va:supportedDataType rdf:resource="http://code-research.eu/ontology/visual-analytics#string" />
        <va:supportedDataType rdf:resource="http://code-research.eu/ontology/visual-analytics#date" />
        <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
        <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
      </va:Axis>
    </va:hasVisualChannels>
    <va:hasVisualChannels>
      <va:Axis rdf:about="http://code-research.eu/ontology/visual-analyticsBarChartYAxis" >
        <rdfs:label>y-Axis</rdfs:label>
        <va:supportedDataType rdf:resource="http://code-research.eu/ontology/visual-analytics#number" />
        <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
        <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
      </va:Axis>
    </va:hasVisualChannels>
    <va:hasVisualChannels>
      <va:Axis rdf:about="http://code-research.eu/ontology/visual-analyticsBarChartColor" >
        <rdfs:label>color</rdfs:label>
        <va:supportedDataType rdf:resource="http://code-research.eu/ontology/visual-analytics#string" />
        <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Optional" />
        <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
      </va:Axis>
    </va:hasVisualChannels>
    <va:hasDescription>
      <rdfs:label>Bar Chart is a diagram that presents the numerical values of variables by the length of bars.</rdfs:label>
    </va:hasDescription>
  </va:BarChart>
</rdf:RDF>

```



**Figure 3.2:** Visual patterns for the bar chart and timeline defined in the description vocabulary.

Using visual patterns, the system is able to generate all mapping combinations which are plausible for the data and perceptually correct regarding visual encoding guidelines. Section 3.2.1.2 we details how VizRec operates to instantiate the appropriate visual patterns– the visual mapping process.

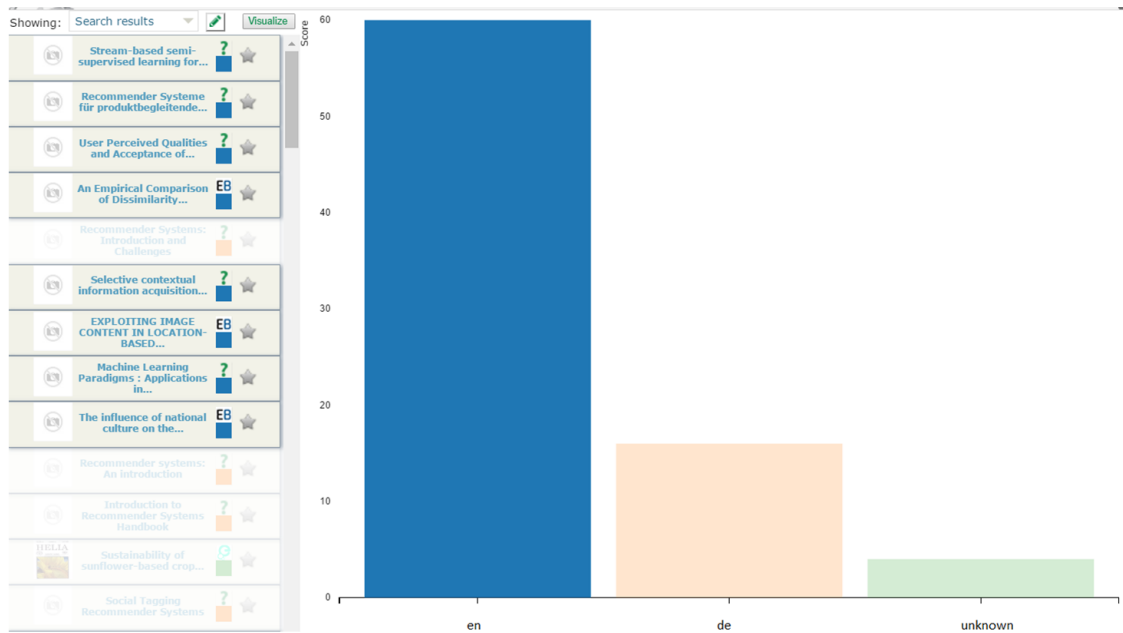
### 3.1.4 Summary

The existing models for describing visualizations are either limited to only few visualizations or store the expert knowledge in a idiosyncratic format that can not be rendered by the machines. The contribution of this section is a machine-readable model that describes visualizations with simple facts that, integrated in rule-based recommender, aids automated visual recommendations. This solution targets real-time recommendations without requiring user to manual specify the visual information that would interrupt the exploration flow of the underlying data. The following section details our rule-based recommender that uses the defined visualization model.

Our vocabulary do not cover all knowledge about the visualization domain, instead it comprises simple facts that will aid the sensible mapping from data to corresponding visualizations.

## 3.2 Rule-based Recommender

This section details the rule-based recommender, analyzing its main concept, workflow and the infrastructure. In particular, the focus is drawn upon the role of the used algorithm (mapping algorithm, see Section 3.2.1.2) in reducing the uncertainty in definition of possible visualizations caused by using description models for visu-



**Figure 3.3:** Recommendation Dashboard using VizRec to automatically recommend appropriate visualizations. On the left side, the list with the recommendations is shown. On the right side one of the recommend visualization is shown which the user uses to filter the recommendations regarding to their language (English, German, unknown).

alizations and data characteristics (see Section 3.2.1.2).

This section contributes 1) a mapping algorithm that automatically defines the list of appropriate visualizations (in further text, mapping combinations) for a given dataset, 2) real-word scenarios where the recommender is well applicable and 3) a user experiment strengthen the usability and applicability of the rule-based recommender presented here. The work in the following sections are based on the publications P6 [MVTS15a] and P7 [MVT16].

### 3.2.1 System Overview of the Rule-based Recommender

The rule-based recommender has been applied in two research projects each with a different goal. The goal of the first project, the EU-funded CODE (Commercially Empowered Linked Open Data Ecosystems in Research) project, was to automatically visualize research data previously extracted from scientific publications to facilitate their exploration. The goal of the EU-funded EEXCESS (Enhancing Europe’s Exchange in Cultural Educational and Scientific Resources) project, however, was to automatically recommend relevant cultural, scientific and educational content (i.e., papers, images, videos) directly to the users habitual environment (browser, content management systems, mobile platforms). Integrated in Recommendation Dashboard [TdSMS15], a tool within the EEXCESS platform that organizes recommended items for visual analysis, the rule-based recommender was used to automat-





**Figure 3.4:** Schematic representation of the rule-based recommender: The stages (a), (b) and (c) illustrate the preprocessing unit. The stage (d) illustrates the visual mapping process between the elements and the visual patterns, whereby the defined mapping combinations are shown in stage (e). Finally, the recommendations will be presented to the user in a random order (f).

ically define visualizations for the recommended items (see Figure 3.3). Yet, each of this projects uses a certain data model with a predefined schema. The CODE project, for instance, uses RDF Data Cubes (see Section 2.1.2.3) which has a well defined structure. The data used in the EEXCESS project are provided by a federated recommender system that collects data from various data sources, such as Mendeley, Europeana, ZBW (German National Library of Economics), ACM Digital Library etc. Each of those sources collects and indexes various kinds of information (books, journals, images, videos, etc.) in repositories structured according to a proprietary (often closed) data model. Before passing to the rule-based recommender, the collected data are structured after a common data model following the specification described in [OHBR15].

Once the data (CODE or EEXCESS data) are passed to VizRec’s rule-based recommender (see Figure 3.4), the recommender applies visual encoding guidelines to generate a collection of visualizations appropriate for the given data. Visual encoding guidelines are generic principles that establish relations between visual channels of a visualization (e.g.,  $x$  – axis of a bar chart) and elements of the data (e.g., whether a field is numeric, categorical, a location, see Section 2.1.1.1). A

preprocessing unit analyzes the data to structure them in terms of interesting data elements so that visual encoding can take place. The two steps to generate visual recommendations, summarized in Figure 3.4, are: (1) preprocessing and (2) visual mapping. This section details and illustrates each unit with a real example of generating visualizations for data obtained from MovieLens.<sup>2</sup> The example is an excerpt of the datasets used for the study in section 5.3. Note that the used dataset simulates a usual output of the EEXCESS's federated recommender. A use case for the CODE project is introduced in Section 3.2.2.

### 3.2.1.1 Preprocessing

The preprocessing unit models, extracts and manages the input data. Furthermore, it addresses the task of prior organization of the visualizations into visual patterns that can be used to reify visualizations. The following describes how these stages go *from data to semantically enriched data* and *from visualization vocabulary to visual patterns* that can be used to actively derive appropriate visual encoding.

#### *From Data to Semantically Enriched Data*

Associated data sources, such as, Mendeley, Europeana, ZBW (German National Library of Economics), ACM Digital Library etc., collect and index various kinds of information (books, journals, images, videos, etc.) in repositories structured according to a proprietary (often closed) data model. For instance, scientific digital libraries, define the structure of literature archives in terms of some important metadata, such as title, abstract, author, keywords, following, e.g., the Dublin Core metadata format.<sup>3</sup>

When it comes to working with the data from idiosyncratic data models in a holistic way, a unified data model offers the following benefits:

- it simplifies the automated data processing, e.g., in terms of extracting information
- it enables interoperability with other applications, e.g., the visualization tools.

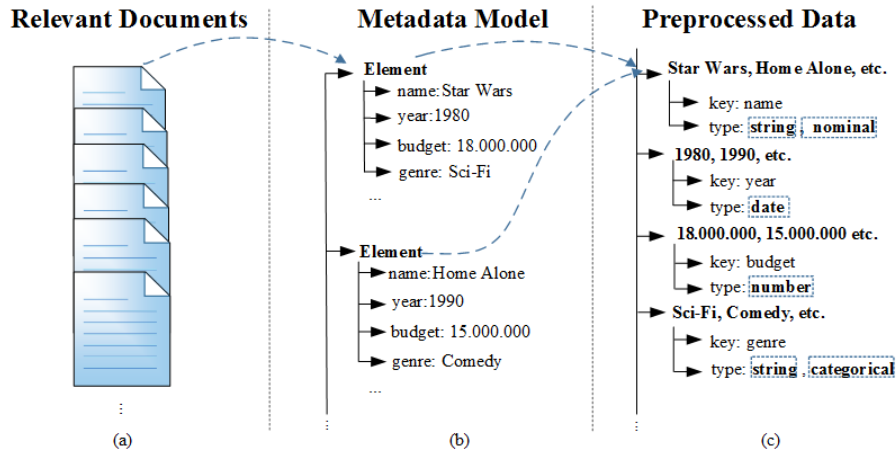
For VizRec, the input data is structured in a common data model following the specification in [OHBR15] (see Figure 3.5 (b)). The unified model organizes metadata elements extracted from the original sources (such as title, content information, document type and attributes). A mapping ontology defines the relation for each concrete metadata format and the unified data model used in VizRec.

The responses to user's query are cached and separately translated into the unified model. Aforementioned digital repositories have specific services to obtain their data with various interfaces for the access, such as JSON, RDF or XML. In

---

<sup>2</sup><https://movielens.org/>

<sup>3</sup><http://www.dublincore.org/usage/documents/overview/>



**Figure 3.5:** Preprocessing: The input data (a) for VizRec are structured following a metadata model (b). The preprocessing unit is responsible, first, for the identification and extraction of the metadata elements and second, for the datatype analysis on the values of the metadata elements (c).

contrast, linked data have a graph structure connecting data that originate from different sources. Data obtained from *DBpedia* or *MovieLens* are translated locally in the common metadata model. Finally, a simple matching operator merges the metadata of each dataset together and presents them as a single dataset. As example, a user interested in budget and income of movies at the end of last century obtains Listing 3.3 for the query *Top 10 successful movies filmed in 1960, 1970, 1980 and 1990*. The dataset merges *movie name*, *genre* and *year* obtained from *MovieLens* with *budget* and *gross* information obtained from *DBpedia*. Aforementioned metadata mapping methods are beyond the scope of *VizRec* and are only mentioned here for completeness. For further details, refer to [OHBR15].

Once the data are obtained, the preprocessing unit of *VizRec* carries out four important technical steps. First, metadata extraction, the metadata elements (i.e., movie name, genre, year, budget and gross) are automatically identified and their values extracted following the metadata model (see Figure 3.5 (b)). Second, datatype categorization, extracted values are collected in series and a data analysis step categorizes them into standard datatypes, such as categorical, temporal and numerical – represented by primitive datatypes string, date and number, respectively (see Figure 3.5 (c)). Third, semantic extraction, if required, using gazetteer lists specialized datatypes are derived, e.g., spatial information like coordinates are obtained for metadata elements belonging to the term *country*. Fourth, enrichment, extracted elements enriched with categorized values are passed to the mapping algorithm to execute the mapping process (see Section 3.2.1.2).

**Listing 3.3:** Exemplary input data for the VizRec obtained from MovieLens and enriched with data from DBpedia.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
<description> The top 10 successfully movies filmed at 1960, 1970, 1980 and 1990</description>
<results>
  <result>
    <facets>
      <provider>DBpedia</provider>
      <type>Linked Open Data</type>
      <moviename>Star Wars: Episode V</moviename>
      <genre>Sci-Fi</genre>
      <year>1980</year>
      <budget>18.000.000</budget>
      <gross>290.158.751</gross>
    </facets>
  </result>
  <result>
    <facets>
      <provider>DBpedia</provider>
      <type>Linked Open Data</type>
      <name>Home Alone</name>
      <genre>Comedy</genre>
      <year>1990</year>
      <budget>15.000.000</budget>
      <gross>285.761.243</gross>
    </facets>
  </result>
  ...
</results>
```

### *From Visualization Vocabulary to Visual Patterns*

Formally, a visualization can be broken down in a number of  $r$  visual channels, each of which encodes a single piece of information visually [Ber83]. One can naïvely think that every visual channel may encode any kind of data. Thus, the possible number of combinations for one visualization is the permutation relation [GSGC08b]:

$$C_r^n = \frac{n!}{(n-r)!}, \quad (3.1)$$

where  $n$  is the number of metadata elements in the dataset (i.e., number of fields). For example, a simple bar chart has three visual channels:  $x$ ,  $y$  and  $color$ . The example dataset in Listing 3.3 has five metadata elements ( $n = 5$ ), so the total number of combinations for the bar chart ( $r = 3$ ) is:

$$\frac{n!}{(n-r)!} = \frac{5!}{(5-3)!} = 60 \quad (3.2)$$

So, the number of options a user would have to consider is rather high even for a simple chart, without considering alternative visualizations. The fact is that many of these combinations are perceptually incorrect, since visual channels are often suited to represent only certain metadata given by the perceptual properties of the channel and the characteristics of the metadata [Ber83]. To prevent the

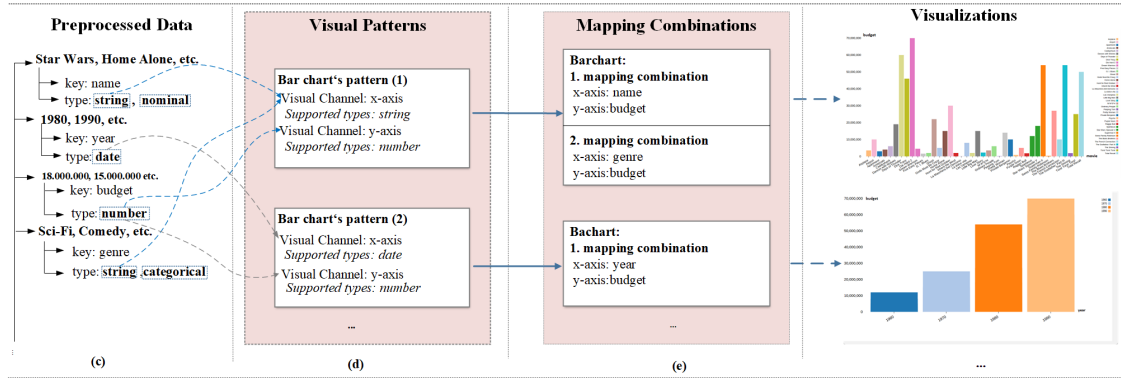


Figure 3.6: Visual mapping process.

generation of incorrect charts, VizRec uses visual patterns to explicitly define which metadata element is related to which visual channel of a visualization type [RB01]. the patterns specify the types of data that are required for each visualization to be instantiated (see Section 3.1.3).

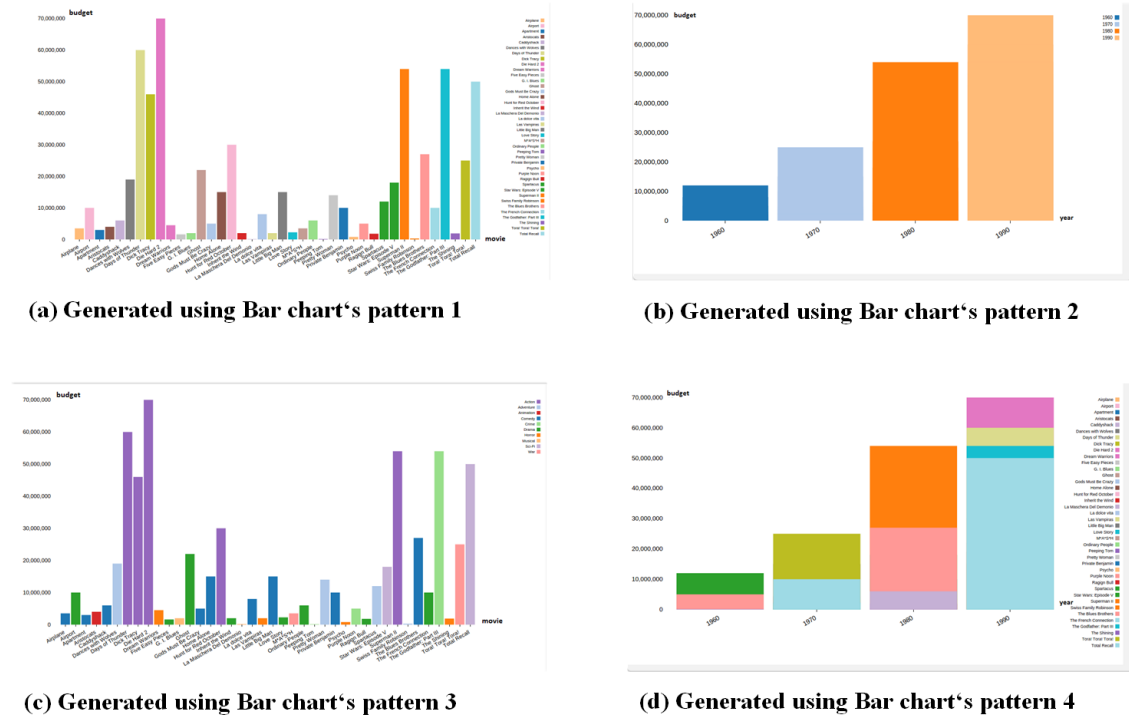
### 3.2.1.2 Visual Mapping

The visual mapping process can be considered as a schema matching problem [RB01]. The basic idea behind schema matching is to figure out a semantic relevance between two objects in schemas under consideration. The result is a mapping comprising a set of elements, each of which indicates that certain elements of schema S1 are related to certain elements of schema S2. In our case, the schemas we deal with are on the one hand the metadata model which describes the semantics of the input data and on the other hand the VA Vocabulary which describes the semantics of the visualizations. Hence, the schema mapping in our context produces mappings each of which describes the correspondence between a metadata element and a visual channel of a visualization to define a possible configuration. In the following we describe this process more in detail.

The relation from elements of the input data to channels of a visualization is valid only if we can establish syntactic correspondences between the metadata and the visualizations. One possibility to identify this is to verify the datatype compatibility. datatype compatibility in our context means having exactly the same datatypes, conforming to the *XSD datatype definition*.<sup>4</sup> The preprocessing unit provides patterns for visualizations and a common model for the input data both including the datatypes of their elements. From the specifications of the visual patterns, the mapping operator compares the datatypes of the visual channels and metadata with each other and builds a list of plausible mappings (see Figure 3.6 (c) and (d)).

Beyond the datatype compatibility, a valid mapping needs to account for structural compatibility, since visualizations have either fixed or varying number of visual

<sup>4</sup><http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/>



**Figure 3.7:** Some of the bar chart combinations generated for the dataset *movies* using the bar chart patterns (1), (2), (3) and (4). The patterns are listed in Table 3.1.

channels. To generate a visualization, the mapping operator has to instantiate every mandatory visual channel while the pattern(s) including optional channels can be ignored if there is no corresponding data element therefor. Formally, each pattern  $i$  defines for each visual channel  $j$  which  $r_j$  metadata element should be selected from  $n_j$  metadata elements:

Note, that  $n_j$  is a subset of  $n$  that complies with datatype compatibility for the  $j$  visual channel. To obtain the total number of combinations  $M_i$ , generated for a particular pattern  $i$ , we multiply every suitable  $\binom{n_j}{r_j}$  notation of a pattern:

$$M_i = \prod C_{n_j}^{r_j} \quad (3.3)$$

$$\frac{n_j!}{r_j!(n_j - r_j)!} = \binom{n_j}{r_j} = C_{n_j}^{r_j} \quad (3.4)$$

Thus, the final number of combinations  $M$  for a visualization is nothing else but the sum of every  $M_i$ :

$$M = \sum \{M_i\} \quad (3.5)$$

Continuing the example about successful movies, VizRec considers the following facts: (i) the underlying dataset contains two *string* values (*movie name*, *genre*),

one *date* (*creation year*) and two *numbers* (*budget and gross*), (ii) only the patterns, which accept categorical/nominal (*string*), temporal (*date*) and numerical values (*number*) are appropriate. Thus, geographical visualizations will not be further considered by the system.

According to the pattern description from Listing 3.1, bar chart complies with facts (i) and (ii) from above. Using visual patterns (see Section 3.1.3), the system selects bar chart pattern (1) counting exactly one element with datatype *string* and one with datatype *number*, producing:

$$M_1 = C_2^1 \times C_2^1 = \binom{2}{1} \times \binom{2}{1} = 4 \quad (3.6)$$

mapping combinations, containing e.g.,  $\{x - axis : movie\ name, y - axis : budget\}$  (see Figure 3.7 a). For pattern (2) the system selects one metadata element with the datatype *date* and one with datatype *number*, obtaining

$$M_2 = C_1^1 \times C_2^1 = \binom{1}{1} \times \binom{2}{1} = 2 \quad (3.7)$$

mapping combinations containing e.g.,  $\{x - axis : creation\ year, y - axis : budget\}$  (see Figure 3.7 b). For pattern (3) VizRec selects one element with datatype *string*, one with datatype *number* and another one with datatype *string*, so that

$$M_3 = C_2^1 \times C_2^1 \times C_2^1 = \binom{2}{1} \times \binom{2}{1} \times \binom{2}{1} = 8 \quad (3.8)$$

mapping combinations are determined containing e.g.,  $\{x - axis : movie\ name, y - axis : budget, color : genre\}$  (see Figure 3.7 c). Note, these mappings contain redundant mapping, that is, an element can be selected for more than one visual channel. For instance, for this pattern the element *movie name* can be mapped once on the *x - axis* and once on the *color* (see Table 3.1 Pat.num (3)) since both channels support the datatype *string*. Redundant mapping with optional visual channels reinforces aspects of the data, in this case the entity movie name [RL95].

Applying the same approach for pattern (4) results in 4 possible mapping combinations containing e.g.,  $\{x - axis : creation\ year, y - axis : budget, color : movie\ name\}$  (see Figure 3.7 (d)). The total number of perceptually valid combinations is 18 (see Equation 3.5 and Table 3.1). Comparing with the naïve result (60) the number of visualizations to consider is reduced considerably. Yet, the example concentrated only on a single chart (bar charts), each type of chart adds another number of visualizations that may be useful for the user. Furthermore, some users may be more inclined to use one type of chart than other to spot what they are looking for in the data. In the following section we consider recommendation strategies to filter results according to the user preferences.

---

**Algorithm 1:** Simplified algorithm for determining appropriate mapping combinations
 

---

```

Data: set(data_element) // retrieved and preprocessed content from data
        sources
Result: set(mapping_combination)
// result set
set(mapping_combination) ← empty set;

// map containing visualizations and their visual patterns
map(visualization, set(visual_pattern)) ← empty map;

// first step: collect all available visual patterns
set(visualization) ← get all visualizations from repository;
while set(visualization) not empty do
    visualization ← take current visualization from the set;
    set(visual_channels) ← get visual channels from visualization;





    // The generation is based on channel attributes: occurrence,
    persistence
    set(visual_pattern) ← generate patterns out of set(visual_channels);

    // Store generated patterns
    map(visualization, set(visual_pattern)) ← append pair
        (visualization, set(visual_pattern));

// second step: identify mappings based on visual patterns
while map(visualization, set(visual_pattern)) not empty do
    visualization ← take current vis. pair (visualization, set(visual_pattern)) from set;
    while set(visual_pattern) not empty do
        visual_pattern ← take current pattern from set;
        while set(data_element) not empty do
            // The structure is evaluated based on a number of visual
            channels
            // within a pattern
            if structural match between (visual_pattern) and (data_elements) then
                // Datatype match is performed between visual channels
                // and individual elements of the current data
                if datatype match between (visual_pattern) and (data_elements) then
                    // Elements of the current data are mapped (linked)
                    // to the corresponding visual channels of the current
                    pattern
                    mapping_combination ← map elements to visual channels;
                    set(mapping_combination) ← append mapping_combination;
                else
                    continue;
            else
                continue;
  
```

---

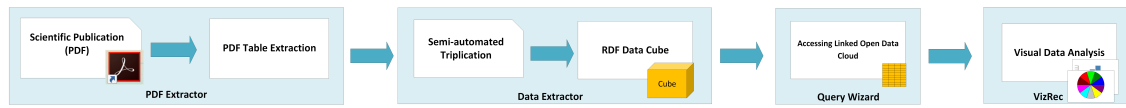


Pat. num.	Visual Patterns	Mappings	Vis.
1	{ <i>x-axis:string, y-axis:number</i> }	{ <i>x-axis:movie, y-axis:budget</i> } { <i>x-axis:movie, y-axis:gross</i> } { <i>x-axis:genre, y-axis:gross</i> } { <i>x-axis:genre, y-axis:budget</i> }	
2	{ <i>x-axis:date, y-axis:number</i> }	{ <i>x-axis:creation year, y-axis:budget</i> } { <i>x-axis:creation year, y-axis:gross</i> }	
3	{ <i>x-axis:string, y-axis:number, color:string</i> }	{ <i>x-axis:movie, y-axis:budget, color:genre</i> } { <i>x-axis:movie, y-axis:gross, color:genre</i> } { <i>x-axis:movie, y-axis:budget, color:movie</i> } { <i>x-axis:movie, y-axis:gross, color:movie</i> } { <i>x-axis:genre, y-axis:budget, color:movie</i> } { <i>x-axis:genre, y-axis:gross, color:movie</i> } { <i>x-axis:genre, y-axis:budget, color:genre</i> } { <i>x-axis:genre, y-axis:gross, color:genre</i> }	
4	{ <i>x-axis:date, y-axis:number, color:string</i> }	{ <i>x-axis:creation year, y-axis:budget, color:movie</i> } { <i>x-axis:creation year, y-axis:gross, color:movie</i> } { <i>x-axis:creation year, y-axis:budget, color:genre</i> } { <i>x-axis:creation year, y-axis:gross, color:genre</i> }	

**Table 3.1:** Mapping combinations defined for the exemplary dataset *movies* using bar chart’s visual patterns (1)-(4). The visualizations shown in the last columns are generated for the first mapping combination of each pattern, to give an example for instantiated mapping combinations.

Having obtained all valid mapping combinations (see Figure 3.4 (e)), the mapping operator maps the values of the selected metadata elements to the corresponding visual channels of a visualization and presents them to the user as a set of appropriate visualizations. The various mapping combinations present different analysis scenarios and thus can cater to wider range of the user needs and interests.

The pseudo-code 1 summarizes the essential steps performed by the mapping algorithm. Initially, for a given dataset (cf., Figure 3.4), relevant visual patterns are identified from the existing visualization collection. Based on those patterns, the schema mapping part of the algorithm identifies the concrete configurations for visualizations that are compatible with the data provided, in the datatypes and the structure. Candidates complying with these rules are valid mapping combinations that are in further steps of the VizRec pipeline used for the detailed, personalized filtering.



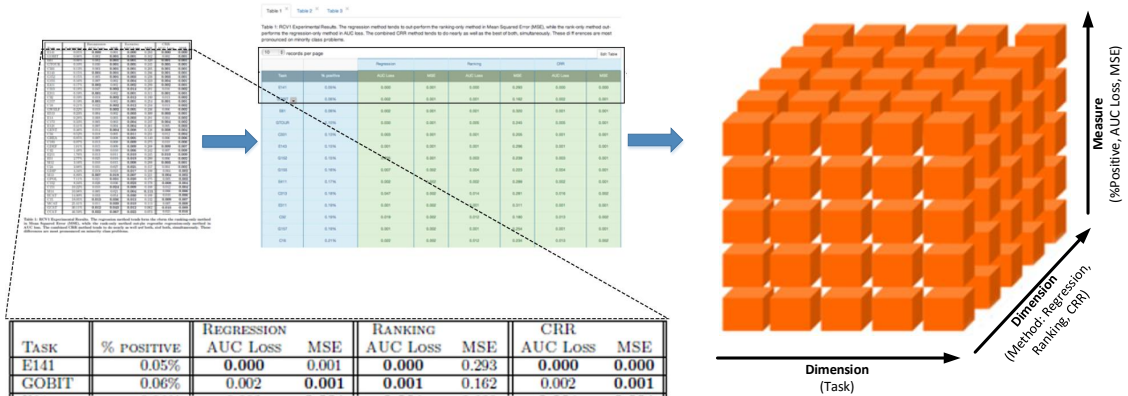
**Figure 3.8:** Schematic representation of the CODE pipeline to extract and analyze scientific facts embedded in publications.

### 3.2.2 Usage Scenario: Automated Visualization of Scientific Data

As already announced in Section 3.2.1, we used the rule-based recommender in two research projects, EU-funded CODE and EEXCESS projects. A usage scenario for the EEXCESS project has been introduced throughout the Section 3.2.1 in order to describe the main concept, workflow and the infrastructure of the recommender. This section introduces the usage scenario for the CODE platform that aims to automatically extract and visualize research data (in further text: scientific facts) embedded in scientific publications. This scenario is of particular interest since it has been used to evaluate VizRec’s rule-based recommender performing a user study (see Section 3.2.3). Note that this section is based on our Publication [MS15].

As shown in Figure 3.8 the CODE platform consists of four units. The first unit, the PDF Table Extractor [KK13] is responsible for extracting the tables from PDF files. After this, the Data Extractor unit organizes the tables in a uniform format following the specification of the RDF Data Cube Vocabulary (see Section 2.1.2.3). For the transformation and semantical enrichment, the extractor uses a HTML table. The columns of the table are automatically classified as either nominal/categorical data (if the cell content is non-numeric), numerical data (for numeric cell content) or multi-value data (if there are multiple values in at least one of the cells of the given column). This process is illustrated in Figure 3.9 and described in [SSK<sup>+</sup>14] in detail. The Data Extractor further releases the Cubes in a Linked Data endpoint to make them publicly accessible.

Yet, in order to access the RDF Data Cube(s), one can use the CODE Query Wizard which serves as an interface to the corresponding Linked Data endpoint. Query Wizard presents the graph structure of the Cube content to the user in an easy-to-use web-based interface very similar to spreadsheet applications (e.g. Microsoft Excel Table). In this table, a row corresponds to a single subject and a column represents a predicate (see Section 2.1.2). Cells contain objects, i.e. any number of literals and/or entities, depending on the row and column. Yet, the biggest advantage of this table is that the user can perform various exploration tasks using the integrated operations which would otherwise require experiences in semantic technologies. Concretely, the user can select columns to be shown, load any number of rows or even perform operations such as filtering and aggregating the listed data. For more details about the Query Wizard we refer to the publication [HGVS14].



**Figure 3.9:** Pipeline to extract and structure scientific facts embedded in publications.

To finally visualize the Cube content, the user activates *VizRec* on the Platform which responds to user’s Data Cube with a list of appropriate visualizations. Once activated, *VizRec* analyzes the structure of the Cube in order to identify and extract the content (i.e., dimensions and measures) being visualized. RDF Data Cube has a well defined structure so that the dimensions and measures are readily to detect and categorize regarding to their datatype (see Section 2.1.2.3). Once the components (dimensions and measures) are extracted, *VizRec* loads the visual patterns and activate the mapping algorithm to perform the visual mapping process. Using the patterns and the mapping algorithm, the recommender maps the components of the cube to the visual channels of the candidate visualizations. To do so, it creates the list of appropriate visualizations for user’s current dataset and presents the listed visualizations to her in a random manner. Now, the user is able to visually explore the content of the tables embedded in scientific publications and uncover insights from the data being represented which might remain hidden by just considering them in a tabular form.

Note, the above-mentioned components PDF Extractor, Data Extractor and Query Wizard are provided by partners of the EU funded CODE project and are behind the scope of this thesis. For more details we refer to the corresponding publications.

### 3.2.3 Evaluation

We performed a preliminary evaluation to investigate how readily non-experts can perform complex analysis tasks using our visual recommender. Additionally, the evaluation results should help us to identify usability gaps, investigate the reaction to recommended visualizations and mapping combinations. For the evaluation we made use of the CODE Platform. However, instead of following the complete workflow from extraction to visualization, we concentrated on the latter part. Thus, we used datasets that had been previously structured as RDF Data Cubes and published in

The screenshot shows the CODE Linked Data Query Wizard interface. At the top, there's a header with the CODE logo and the text "Linked Data Query Wizard" and "Watch the screencast". Below the header, there are several action buttons: "Visualize this dataset", "MindMap the 10 displayed results", "Aggregate this dataset", and "For the Geeks". The main content area displays the title "G20-Countries: Co2 (Tons per Person), Life Expectancy, Population; 2000-2009" and the source "(Source: CODEV)". A table with 7 columns is shown: Country, Region, Subregion, Year, Co2 Tons/P, Life Expectancy, and Population. The table contains 10 rows of data for Japan from 2000 to 2009. Below the table, it says "Displaying 10 of 90 results" and there are two buttons: "Load 10 more results" and "Load all results".

Country	Region	Subregion	Year	Co2 Tons/P	Life Expectancy	Population
Japan	Asia and the Pacific	Eastern Asia	2000	9.69996009634434	81.35	125,720,310
Japan	Asia and the Pacific	Eastern Asia	2001	9.54899042026934	81.7	125,893,623
Japan	Asia and the Pacific	Eastern Asia	2002	9.6521706040997	82.0	126,048,366
Japan	Asia and the Pacific	Eastern Asia	2003	9.80535201770866	82.07	126,184,149
Japan	Asia and the Pacific	Eastern Asia	2004	9.97268285029414	82.36	126,299,414
Japan	Asia and the Pacific	Eastern Asia	2005	9.79539113516186	82.27	126,392,944
Japan	Asia and the Pacific	Eastern Asia	2006	9.73540548112566	82.68	126,464,789
Japan	Asia and the Pacific	Eastern Asia	2007	9.88855757152132	82.87	126,515,486
Japan	Asia and the Pacific	Eastern Asia	2008	9.54269049246179	82.81	126,544,640
Japan	Asia and the Pacific	Eastern Asia	2009	8.70026998055854	82.932	126,551,705

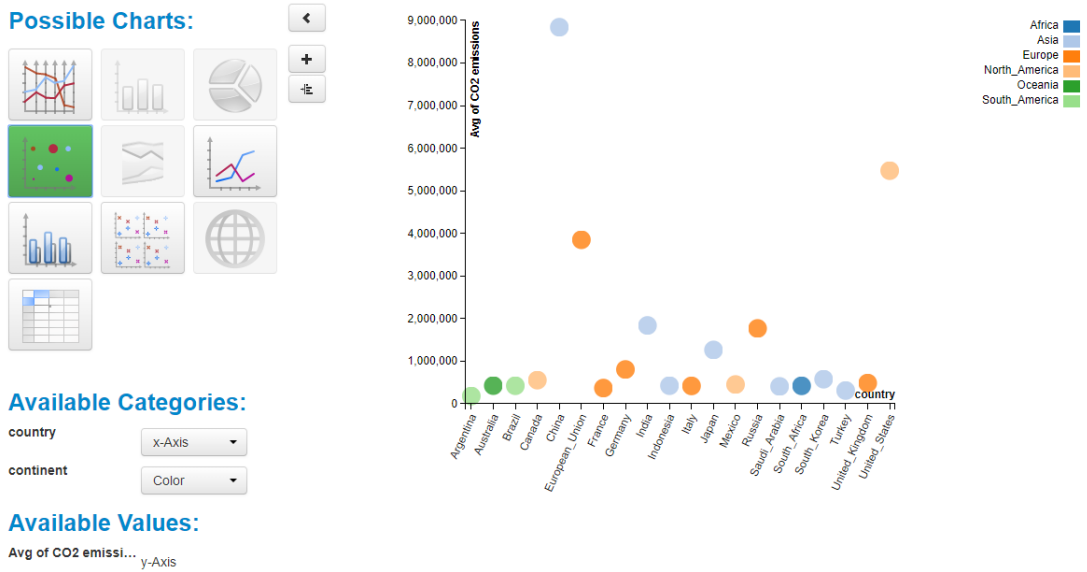
**Figure 3.10:** The Query Wizard serves as an interface to the Linked Data endpoints. Once a RDF Data Cube is queried from the endpoint, Query Wizard presents the graph structure of the Cube content in a tabular form. Using the operations (aggregation, filtering, removing/adding columns) provided with the table, user can perform various exploration tasks on her data before passing it to VizRec.

our own Linked Data repository. This insured the independence from other services and components of the CODE workflow and offered a more controlled evaluation environment.

During the experiment, participants were mainly exposed to the Query Wizard and VizRec. The Query Wizard served as an interface to select the predefined Cubes from the corresponding repository and is one of the components for visual analysis inside of the CODE platform.

### 3.2.3.1 Procedure

The evaluation procedure started with a demonstrative video of the UI features of VizRec and Query Wizard. Two individual datasets were used to demonstrate the use of Query Wizard in displaying and filtering RDF Data Cubes (see Figure 3.10) and VizRec (see Figure 3.11) in automated visualizing and visually exploring the Cube content. Concretely, we used (i) one dataset with two dimensions and one



**Figure 3.11:** The user interface of VizRec’s rule-based recommender. Once Query Wizard passes the RDF Data Cube to VizRec’s rule-based recommender, the recommender extracts the dimensions and measures (see bottom left) and defines the list of appropriate visualizations (top left). Once the user selects one of the selected visualizations, the visualization will be automatically generated (right).

measure and (ii) one dataset with two dimensions and two measures. During the demonstration each participant received a description on how a visualization is usually recommended in VizRec. We also introduced the fact that VizRec proposes a visualization with a default mapping and described how to set a new mapping. After the demonstration, participants were presented one example dataset showing *the CO2 emissions between the years 2000 and 2009 within the G20 Countries* and six time constrained analysis tasks they had to perform using the dataset. For instance, the first Task (T1) was the following:

(T1) *Please show the data set in Query Wizard. We are interested only in the countries which have a CO2 Emission over 13 Tons per persons. After that, please visualize the results. You have 3 minutes to complete this task.*

To perform the tasks, the participants first had to display the dataset in tabular form on Query Wizard and then visualize it using VizRec. After this, they were required to interact with the recommended visualizations by performing operations such as filtering, aggregation and brushing in order to complete the task. The moderator kept track of the time, the participants consumed for a task, however there was no timer shown to participants. Once the task was completed, the participants had to fill a ten-point likert NASA TLX <sup>5</sup> scale covering six dimensions of workload: mental demand, physical demand, temporal demand, effort, frustration

<sup>5</sup><https://humansystems.arc.nasa.gov/groups/tlx/>

and perceived performance. Furthermore, the participants were asked to answer nine questions for specific UI features of the tool. The results of NASA TLX helped us to measure the subjective workload whereby the provided answers for the questionnaire were used as subjective assessments of usability, for usability analysis respectively. Note that the tasks and the questionnaire used for this evaluation are listed in Chapter 8, Section 8.1.

### 3.2.3.2 Participants

The heuristic evaluation was performed by eight IT experts. Some of the participants were experienced in visualizing data. However, none of the participants had experiences in Linked Open Data or Semantic Technologies. Note, more demographic details are listed in Appendix (see Section 8.1). In total, the participants concluded 48 tasks. 39 were successfully completed in time whereby for 15 tasks participants took more time than assigned. Since VizRec suggests all possible visualizations for a given dataset and participants were free to choose different ones, we did not collect any quantitative measures. We collected subjective feedback towards the overall usage of the VizRec and the appreciation of interacting with mappings.

### 3.2.3.3 Results

The evaluation has been performed as a combination of the subjective assessment of usability (usability analysis) with the NASA Task Load Index. Based on this combination, we generated two different kinds of results of our evaluation:

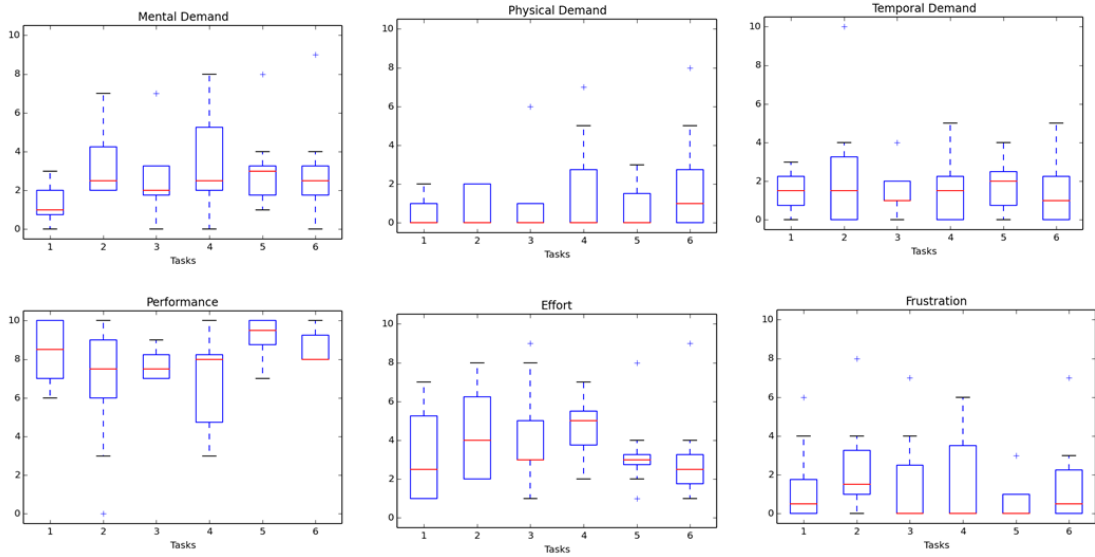
- Subjective Workload: NASA Task Load Index
- Usability Analysis

In the following we present the results of each category.

#### *Subjective Workload: NASA Task Load Index*

The results of the NASA Task Load Index are presented as box plots in Figure 3.12. The six plots represent the results for the six different aspects of the NASA Task Load Index; mental demand, physical demand, temporal demand, performance, effort and frustration. Looking at the plots, we can observe the following:

- The mental demand of the participants was rather low for all tasks. We detected that only by task four the mental demand increased slightly. However, the variance between the participants was quite high.
- The physical demand was constantly low for the first three tasks and increased slightly for the more complex last three tasks.



**Figure 3.12:** Subjective Workload measured using NASA Task Load Index.

- The temporal demand remained generally low with a high degree of variance between the participants. From 48 tasks performed in total by 8 participants, 39 were successfully completed in time, 15 were not completed in time.
- The performance scores were generally high but again with a high degree of variance. From 8 participants and 48 tasks in total, 39 tasks were successfully completed by the participants, 13 tasks could be completed only with moderator’s intervention. 2 tasks could not be completed at all.
- The subjective effort of the participants was differently (showed a very high variance between the participants) depending on the learning effect. As the participants received a task with a new topic, it demanded more effort from the participant and if the task was similar to the previous the effort decreased.
- The frustration was rather low throughout the evaluation but with a high variance between the participants.

### *Usability Analysis*

The section summarizes the statements of the participants provided as answers to the usability questions. The following lists these statements:

- The participants found VizRec easy to use after they have been introduced to the system. The participants especially favor the intuitive and interactive user interface and the way how the buttons, icons and the information were presented there. They found that there is neither too much nor too little information shown on the user interface (see Figure 3.11).

- The participants appreciated the automated recommendation of appropriate visualizations without requiring the user to manually specify the visual information.
- The chart collection was sufficient to all participants. Nevertheless, two participants criticized the color encoding by some visualizations. This preliminary referred to those with a large number on data points where the color codes have been reused after a certain range of possible encodings was reached.
- The default mapping of the visualizations has not been always received well by the participants. The participants wished to have been provided with better mechanism that selects the mapping combinations considering, for instance, the visual needs and preferences of the user.
- Another big critical point was the missing ranking of the recommended visualizations (see Figure 3.11 on the left side). The participants criticized that the visualization were not sorted regarding to their relevance to the user.

#### 3.2.3.4 Discussion

This evaluation helped us to quantify VizRec’s quality in supporting the user in performing complex analysis tasks in terms of workload, completion time and performance. It revealed which part of VizRec is proved useful and easy to handle. Moreover, it provided insights on how the recommendations are perceived by the user. The study was setup as formative and not comparative so we did not measure the statistical deviation from a baseline.

In general, the evaluation results revealed that our approach can support the non-expert user in visually exploring their data and perform complex analysis tasks. However, participants found that the initially recommended visualizations did not always emphasize their visual preferences and needs. Thus, they appreciated the ability to set another mapping combination for a certain visualization or even choose an another visualization. Regarding to the usability of the tool, we could observe an overall user satisfaction.

A clear limitation of the presented evaluation is the small number of participants which prevented us from performing statistical inferences. Thus, this evaluation needs a follow-up evaluation involving more participants.

#### 3.2.4 Summary

In this section we introduced our rule-based recommender which automatically recommends and creates appropriate visualizations for a given dataset. The recommender has been used in the EU-funded CODE- and EEXCESS (Enhancing Europe’s Exchange in Cultural and Scientific Resources) platforms. The goal of CODE was to automatically visualize research data previously extracted from scientific publications. The goal of EEXCESS (Enhancing Europe’s Exchange in Cultural and



Scientific Resources), however, was to automatically recommend relevant cultural, scientific and educational content (i.e., papers, images, videos) directly to the users habitual environment (browser, content management systems, mobile platforms). Integrated in Recommendation Dashboard [TdSMS15], a tool within the EEXCESS platform that organizes recommended items for visual analysis, the rule-based recommender was used to automatically define visualizations for the recommended items (see Figure 3.3).

Yet, our tool is not limited to the scientific or cultural domain. The application scenarios for the approach we propose in this thesis span numerous areas, both scientific and industrial that requires automated visual representation of their data [MHT<sup>+</sup>14]. Although, rule-based VizRec guarantees the effective graphical presentation of the data, the list of the recommended visualizations is too large and ignores the fact that the choice of visual representation is a personal matter, a matter of the users' needs and preferences respectively. As a result, the users are often overwhelmed to select the mapping combinations that better address their visual needs and preferences (see Section 3.2.3). To address this issues, we defined a personalized visual recommender that recommends visualizations considering user's preferences and interest (see Chapter 5). To collect the user preferences we performed a crowd-sourced study involving general public. The procedure and the results of this study are presented in the next chapter.

# Chapter 4

## A Crowd-sourced Study for Preference Elicitation

The results of the user study introduced in Section 3.2.3 revealed that the visual recommendations defined by a rule-based system barely address user’s visual needs and preferences. These recommendations are defined by only considering the properties of user’s data and ignoring the fact that the choice of visualizations is a personal matter. Considering this outcome, we wanted to investigate which visualizations people prefer and if there is a single answer why some people prefer a visualization more than another. This was only possible by collecting user preferences involving the general public. To achieve this, we performed a crowd-sourced study and collected user preferences for visualizations defined by the rule-based recommender.

In this chapter, we detail how we collected the user preferences and report the results of their extensive analysis. Based on that, this chapter makes the following contributions: 1.) a method to elicit user preferences and 2) an extensive analysis of variability in users’ visual preferences. Note that the following sections are based on the publication P7 [MVT16].

### 4.1 Experiment Setup

Our goal in this crowd-sourced study was to obtain preferences for visualizations coming from our rule-based recommender system. We further applied several methods to analyze the variability in preference scores. The following details the study.

#### 4.1.1 Datasets

Visualizations were generated for three open source datasets (see below) using a rule-based visualization recommendation system [MVT16]. Note that the rule-based recommendation system uses heuristic rules that produce visually correct charts, but they are not always useful. Some examples of such charts are given in Figure 4.3. They generally received low ratings since they either were visually useless, do not

reveal much about the underlying data or they do reveal something, but not enough. For instance, the geo chart at the top right shows countries, but it actually hides all data about book publishers which is essential to understand what is being shown. Finally, there were also charts which have shown enough information, but in fact received low ratings. These are typical cases where the user expressed their subjective opinions (cf. stacked bar chart on bottom left).

The following datasets have been used for the experiment:

*MovieLens dataset (Movies)*: This dataset comprises information about the top-ranked movies for the years 1960, 1970, 1980 and 1990. It has 40 entries, which are selected from items of the respective dataset and are characterized by the elements (movie) name, genre, budget, gross, creation year, shooting location (country) and population of the country. Based on this, the mapping unit of the rule-based system produced four types of visualizations using the method described in Section 3.2.1.2 with the following mapping frequencies: 32 bar charts, 9 line charts, 13 timelines and 1 geo-chart. Hence, a total of 55 mapping combinations were generated.

*EU Open Linked Data Portal dataset (EU)*: The EU <sup>1</sup> dataset collects the percentage of the population looking for educational information online in the years 2009–2011 for 28 EU countries. It has 91 entries characterized by elements (country) name, year, language, population, constitutional form and value (in percent) of the population looking for educational information. The mapping unit suggested 30 possible mapping combinations, concretely 15 bar charts, 6 line charts, 8 timelines and 1 geo chart.

*Book-Crossing dataset (Books)*: This dataset <sup>2</sup> contained 41 randomly chosen books published between 1960 and 2003 and characterized by the elements name, country, publisher and year. The mapping unit suggested 3 visualization types: bar chart with 3 combinations, geo chart with 1 combination and timeline with 3 combinations, the total of 7 mapping combinations.

### 4.1.2 Procedure

A crowd-sourced experiment was carefully designed to obtain the user preferences in different formats for each chart. While using a crowd-sourced platform, it is important to design the study so that participants do not blindly click through the options. Thus, before giving a score, a participant had to perform some cognitively demanding task with the visualization (i.e., a minimal analysis). Based on the experiments conducted by Kittur et al. [KCS08], this preparatory task should bring participants to accurately study the chart and prevent a random or rash rating. We designed the task as follows: 1) a participant was given a one line description of

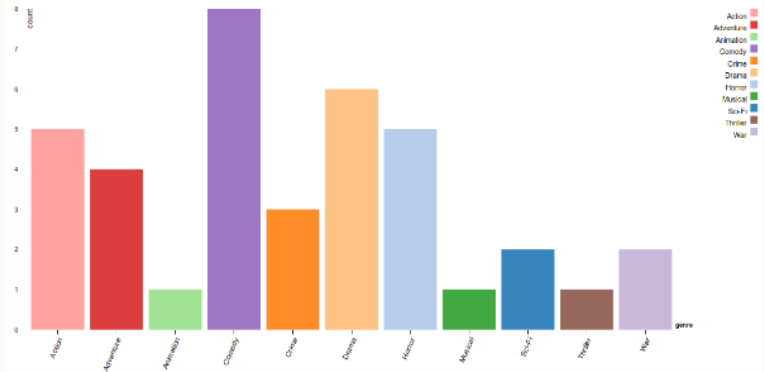
<sup>1</sup><https://open-data.europa.eu/en/linked-data>

<sup>2</sup><http://www2.informatik.uni-freiburg.de/cziegler/BX/>

**Description of the Dataset**

The chart below is about **movies**.

**Chart configuration 3** (please click [here](#) to enlarge the image)



Genre	Count
Action	5
Adventure	4
Animation	1
Comedy	8
Crime	3
Drama	6
Horror	5
Musical	1
Sci-Fi	2
Thriller	1
War	2

**Please write at least 2 (max. 5) tags describing this chart. Hint: Think about what search terms would you apply to find this chart!**

**Please write a title for this chart**

**Please rate the chart above between 1 (means not applicable) and 7 (means very applicable)**

Easy to understand	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Cluttered	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Useful	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Boring	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Organized, clear	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Exciting, interesting	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Effective	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Satisfying	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Confusing	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7

**Please write 1 question, that can be answered by looking the chart**

**Figure 4.1:** Crowd-sourced Experiment Task (HIT). Participants were motivated to carefully observe the visualization with the study task, in terms of writing tags and a title. Thereafter, they had to rate it in a multidimensional scale and pose a question that is answered with the visualization.

a dataset originating the visualization, 2) looking at the visualization she had to write tags (at most five), 3.) write a title, 4) rate the visualization and 5.) write a question the visualization can answer, Figure 4.1 shows an example of a HIT.

The score system used a multidimensional scale adapted from a list of usability factors presented in [SDKP06] and [ZLZK07]. Providing multidimensional rating scale should assist the user in considering various aspects of a visualization and thus to specify the subjective ratings for the considered visualization. The rating scale contained the following factors: (1) cluttered, (2) organized, (3) confusing, (4) easy to understand, (5) boring, (6) exciting, (7) useful, (8) effective and (9) satisfying. Note, that dimensions 1–6 are duplicated with opposing sentiment (e.g., cluttered vs. organized). Opposing dimensions were used to ensure meaningful ratings for scales with complex meaning. Dimensions were rated on a 7-point Likert scale (1=not applicable – 7=very applicable).

Since the visualization scores were also intended for the offline experiment (see Chapter 5), each participant had to rate more than one visualization. We experimented with varying sizes of HITs (Human Intelligent Task), collecting ten (10) and five (5) tasks. After a pilot study these turned out to take overly long (around 15mins), we settled for collecting three (3) charts per HIT. Suggested combinations were distributed in 32 HITs, each of which contained 3 randomly chosen mapping combinations. Pilot studies also helped to streamline dataset descriptions, task descriptions and instructions across the study. After accepting a HIT, the participant (worker or turker) received a tour to complete a task, which showed a visualization and corresponding tags and ratings in the exact same format as the subsequent study. When ready, the worker started the first task in the HIT by pressing a button. Workers were allowed to write *not applicable* or NA for tags but were alerted if they failed to write any tags. The rating dimensions were not assigned a score until the worker did it. Workers could only proceed if they had rated all dimensions. A HIT with three visualizations/combinations was compensated with \$1.00. A worker rated a minimum of three visualizations, but to ensure a more realistic training set for the CF-RS, workers were allowed to perform more than one HIT. Only expert workers who consistently achieved a high degree of accuracy by completing HITs were allowed to take part in the study.

## 4.2 Results

This section presents the results of the crowd-sourced study. It first provides statistics about the participants and next reveals the results of the preference analysis.

### 4.2.1 Participants

Each HIT was completed by ten workers. Note, more demographic details are listed in Appendix (see Section 8.2.1). For 92 visualizations, 8280 ratings across 9 dimensions and 4483 tags were collected from 70 participants. Participants completed on average 4.7 HITs. The experiment started on November 26, 2014 and ended on December 3, 2014. The allotted working time per HIT was 900 sec and the average

	<b>Movies</b>	<b>EU</b>	<b>Books</b>
#visualizations	55	30	7
#users	36	19	15
#ratings	4950	2700	630
#tags	2731	1403	349
#unique tags	292	166	87
Avg. #tags per visualization	49.65	46.76	49.86
Avg. #unique tags per visualization	23.09	22.27	23.43
Avg. #tags per user per visualization	3.23	3.24	3.23
Avg. #unique tags per user per visualization	2.18	1.74	2.61
Avg. #tags per user	75.86	73.84	23.27
Avg. #unique tags per user	20.97	21.00	12.27
Avg. #users per visualization	10	10	10
Avg. #visualizations rated/tagged	15.27	15.79	4.67

**Table 4.1:** Basic statistics of the three rating and tag datasets collected via the crowd-sourced experiment.

working time of workers was 570 sec per HIT. Table 4.1 summarizes the details about the study. As an example the Figure 4.2 presents the three (in average) highest rated visualizations in each dataset whereby the Figure 4.3 the three (in average) lowest rated visualizations in each dataset. Finally, Table 4.2 lists the top-10 stemmed tags in each of the three datasets.

## 4.2.2 Variability in Preference Scores

A set of studies was carried out to analyze the variability in preference scores. To compute the overall score for a visualization for each worker, the scores in opposing dimensions (clutter, confusing, boring) were inverted and then all dimensions were averaged together according to the following equation:  $SC = \left( \sum_{i=1}^k \rho_k D_k \right) / k$ . Where  $k = 9$  is the number of dimensions,  $\rho_k$  is the coefficient 1 and  $D_k$  is  $k$  dimension score. The visualization score was obtained by averaging the worker scores.

The heatmap in Figure 4.4 shows the mean rating for every dimension for each visualization. The results confirm a clear understanding of the opposing dimensions. Negative dimensions, in lower case received opposite scores to corresponding positive ones (UN-co or-cl, EX-bo, in Figure 4.4 top). The aggregated score for each visualization in the bottom row of the heat map (SC) shows that only a handful of visualizations achieved clearly high scores, whereas for each type there were high scoring visualizations. More importantly, the violin plot at the bottom explains these scores: there is a broad variability in scores for most visualization instances.

Rank	tag	#tags	#visuals	#users
<b>Movies</b>				
1	movi	545	55	34
2	genr	252	34	26
3	budget	212	20	22
4	popul	191	32	30
5	gross	130	26	21
6	film	127	48	14
7	year	108	32	17
8	chart	97	46	7
9	titl	90	27	12
10	countri	60	18	16
<b>EU</b>				
1	popul	194	27	18
2	constitut	91	15	13
3	educ	88	16	11
4	republ	86	15	15
5	monarch	79	15	16
6	form	63	19	11
7	countri	62	22	13
8	govern	61	15	10
9	year	60	19	14
10	language	37	8	11
<b>Books</b>				
1	book	93	7	14
2	publish	68	7	15
3	count	35	5	13
4	year	21	4	9
5	titl	12	4	7
6	countri	11	3	7
7	timelin	7	4	3
8	novel	7	6	3
9	famou	6	4	3
10	inform	5	4	1

**Table 4.2:** Top-10 tags in each of the three datasets. The tags were stemmed (normalized) as described in Section 5.1.

The violin plot shows the density of scores; variability is visible in the different shapes as in the spread of the shapes. The coefficient of variation computed for each chart confirmed this assumption ( $M = 0.36$ ,  $SD = 0.12$ ) the minimum variation was 0.07 and the maximum was 0.64, see Figure 4.5. A Levene test on scores con-



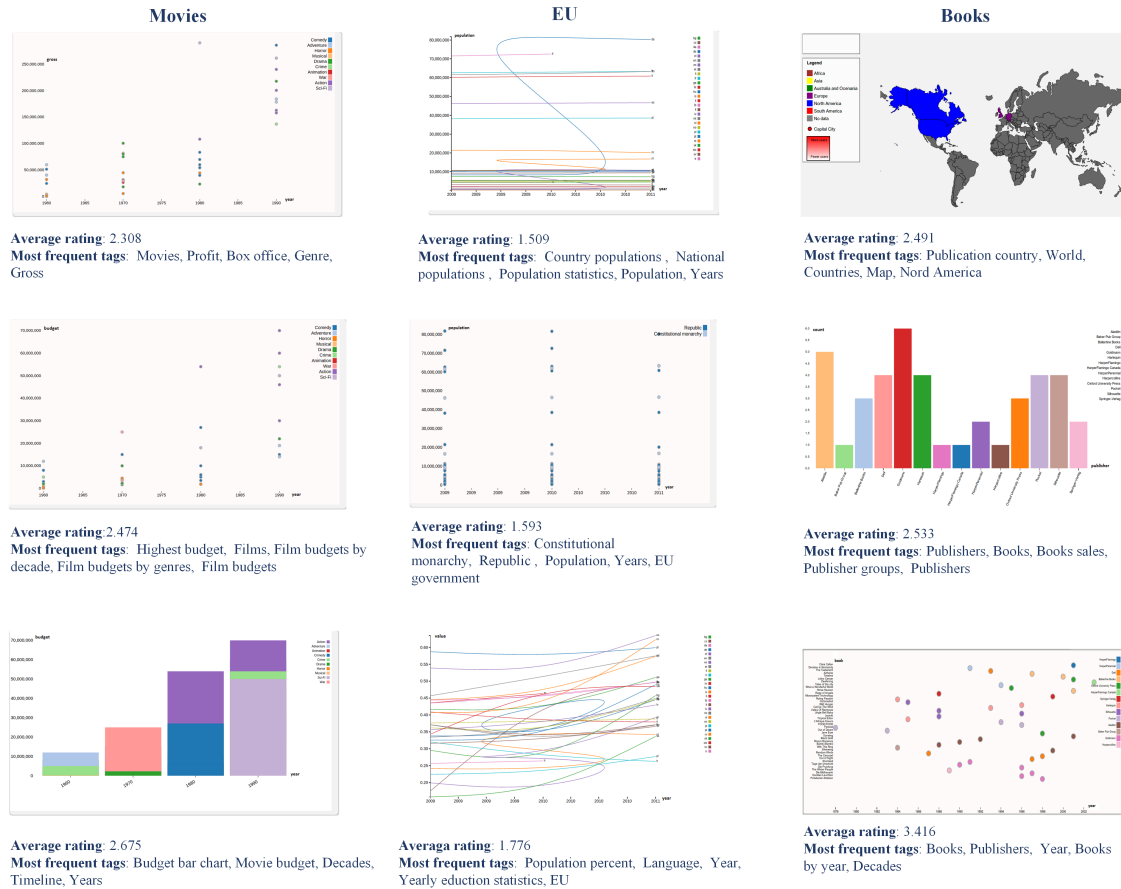
**Figure 4.2:** Three (in average) highest rated visualizations for the datasets *Movies*, *EU* and *Books* including the five most frequently used tags.

firmed significant differences in variances across charts ( $F = 1.64$ ,  $p < 0.001$ ). This supports our assumption that the user preferences matter when choosing the right representation. The results confirm that only a very small number of visualizations achieved high scores and the rest were variable.

From the heatmap individual top-scoring visualizations can be identified. To establish differences in the visualization categories and datasets, we performed a factorial ANOVA with the visualization type and dataset as factors (visualization-type: *bar*, *line*, *time*, *geo* and dataset: *Movies*, *Books*, *Eu*). Homogeneity of variance was confirmed by a Levene test. The factorial ANOVA revealed a significant effect of dataset  $F(2, 908) = 21.19$ ,  $p < 0.0001$ , a significant effect of visualization type  $F(3, 908) = 38.98$ ,  $p < 0.001$  and significant interaction effect dataset visualization type  $F(5, 908) = 3.81$ ,  $p < 0.01$ . TukeyHSD multiple comparisons revealed a significant difference in scores between *Movies* ( $M = 4.86$ ) and *Books* ( $M = 3.82$ )  $p < 0.05$ , as well as between *Movies* and *Eu* data ( $M = 3.68$ ),  $p < 0.001$ . For the visualization type, there was a significant difference in scores between *bar* ( $M = 4.60$ ) and *geo* ( $M = 3.06$ )  $p < 0.001$ , *bar* and *line* ( $M = 3.29$ )  $p < 0.001$ , *bar* and *time* ( $M = 3.72$ )  $p < 0.001$ , as well as between *time* and *line*,  $p < 0.02$ . The significant effects of multiple comparisons for interaction are shown in Figure 4.6.

The main outcomes are the information about the user preferences and the clear differences among them. The interaction effects illustrate several differences amongst



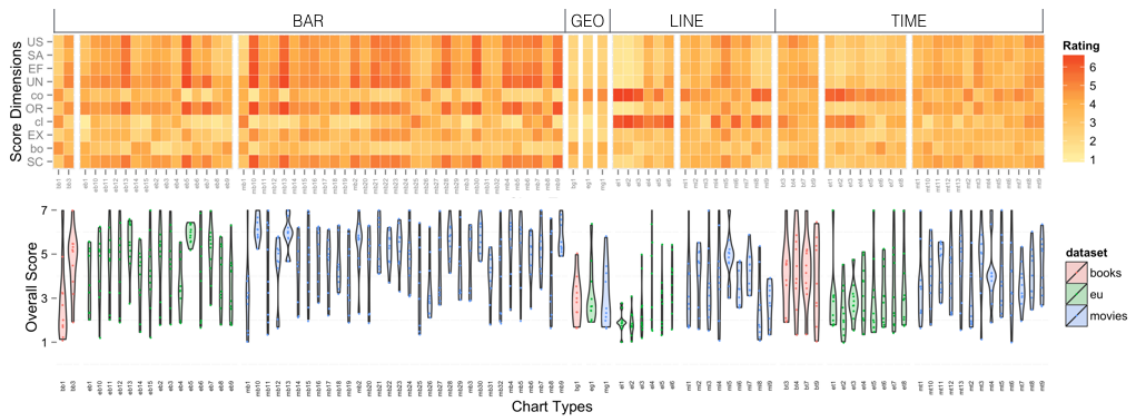


**Figure 4.3:** Three (in average) lowest rated visualizations for the datasets *Movies*, *EU* and *Books* including the five most frequently used tags.

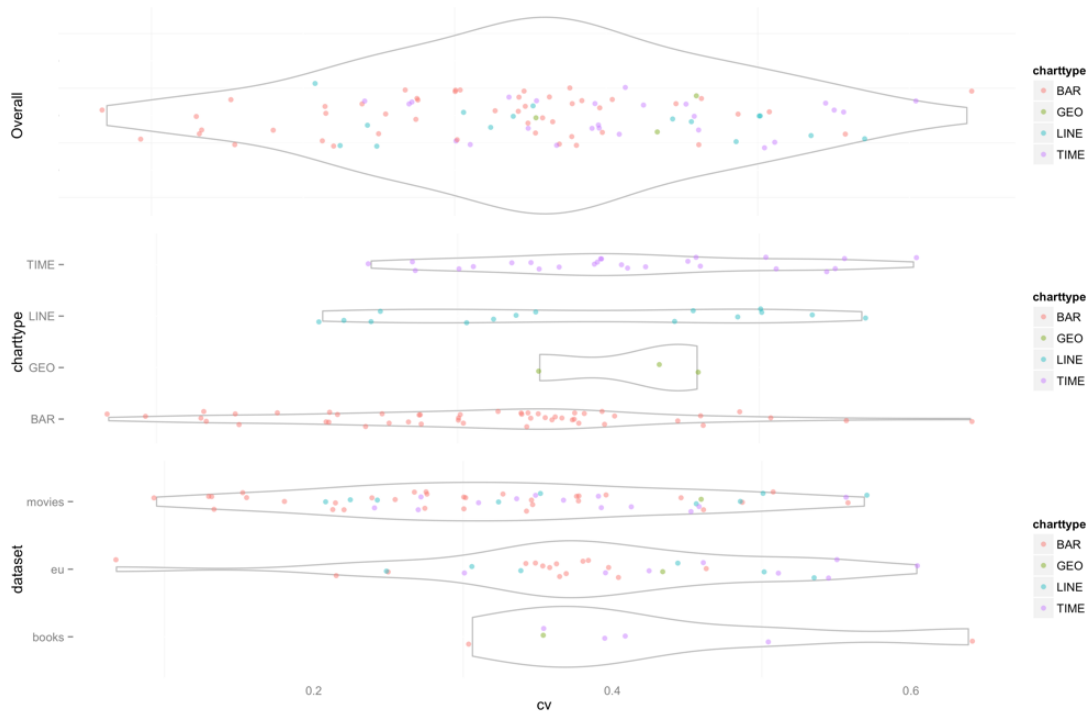
visualization type. For instance, the majority of the users preferred bar chart, probably since it is familiar to most people. Another reason may be that it is easier to compare the values of several numbers at once using bar chart. Yet, these results merely indicate that there are varied preferences. Looking at each dataset, visualization and visualization type in the heatmap of Figure 4.4, it is clear that while a small number of visualizations are generally preferred, in most cases the ratings vary widely and a personalized approach would accommodate those the user preferences better.

### 4.3 Summary

In this section, we present our crowd-sourced study performed to collect user preferences for our personalized visual recommender. This study resulted in enough preferences that could be used for recommending visualizations as well as to confirm our assumption that user preferences matter when choosing the right representation.

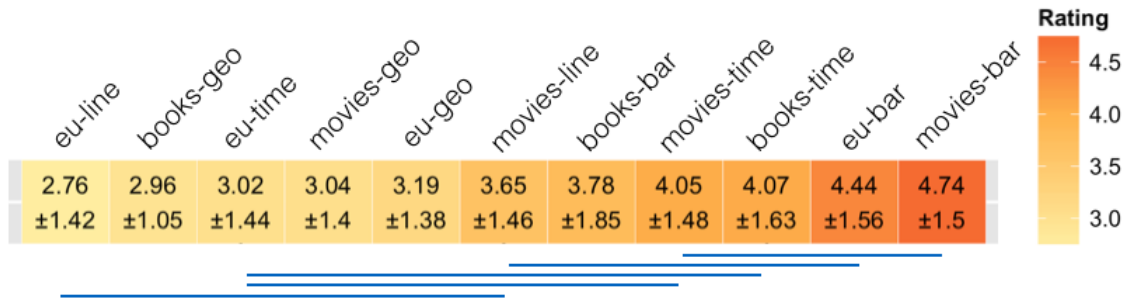


**Figure 4.4:** Mean and variability in scores (rating 1–7, higher is better). The heatmap illustrates the contribution of 9 dimensions (US=useful, SA=satisfying, EF=efficient, UN=Easy to understand, co=confusing or=organized, cl=cluttered, EX=exciting, bo=boring) to the overall score (SC). The violin plot below illustrates the high variability in personal ratings.



**Figure 4.5:** Coefficient of variation. Overall variation for all charts (top), variation broken down by chart type (middle) and by dataset (bottom). Note, that coefficient of variation is the ratio of SD/Mean. The density in the violin chart shows where the broad variation of scores across charts.

Basically, the analysis provided the empirical evidence that visual representation of



**Figure 4.6:** Significant Interactions Visualization Type / dataset. The heat-map illustrates the mean score and standard deviation for each combination of *dataset-visualization type* (1=completely disagree, 7=totally agree). The lines below show where differences begin to be significant. Note, that due to its high variability, *books-bar* is not significantly better than *eu-line*, whereas *movies-line* is.

data is a matter of the users' taste and preferences and only a personalized approach can account for such variability. In Chapter 5, we present our personalized visual recommender which automatically identifies the set of appropriate visualizations using the rule-based algorithm and then filters a subset based on user's preferences and needs.

# Chapter 5

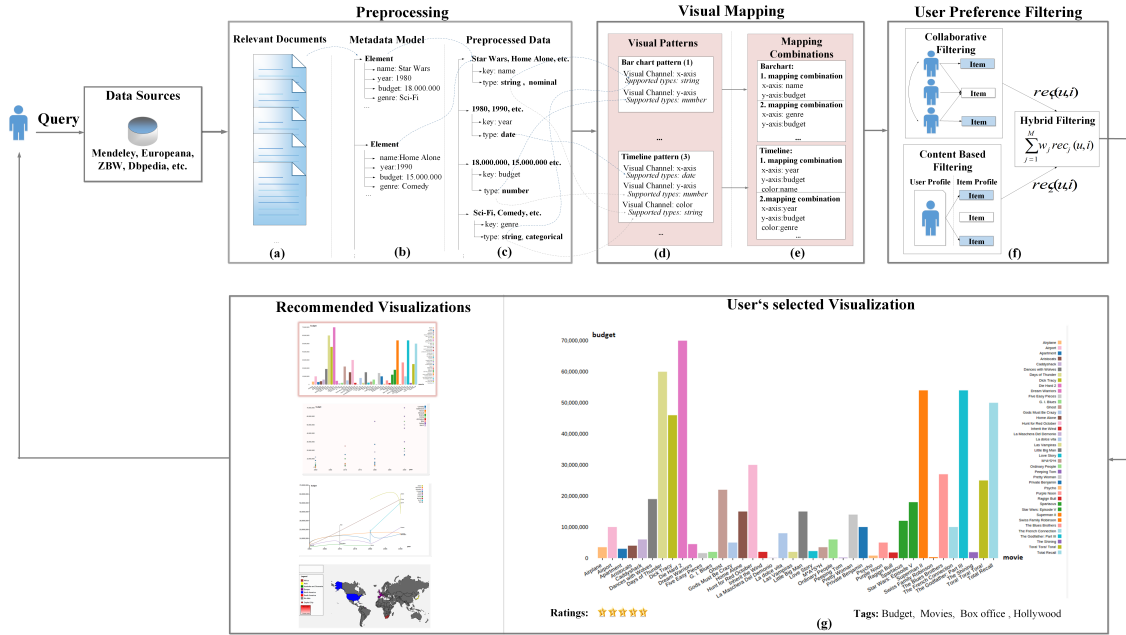
## Personalized Visualization Recommendation

This chapter takes a closer look at the design and construction of the Personalized Visual Recommender. It investigates strategies to recommend personalized visualizations each considering different aspects of user preferences. In particular, the focus is on providing a detailed description of the architecture and the recommendation workflow. We first investigate a collaborative filtering approach based on a multidimensional scale to gauge different quality aspects of the visualizations suggested. We analyze the effects these aspects have on the filtering of visualizations. Content based filtering is analyzed as a means to suggest visualizations according to the information needs of the user, by collecting vectors of tags describing what aspects of the data each visualization represents. A combination of both aspects is studied using a hybrid recommendation strategy.

This chapter contributes 1.) a novel visual approach to recommend personalized visualizations and 2.) a series of studies about recommendation of visualizations. Note, the following sections are based on the publication P7 [MVT16].

### 5.1 Approach

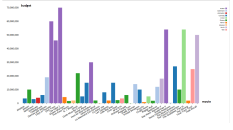
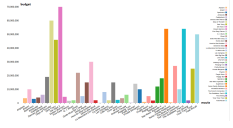

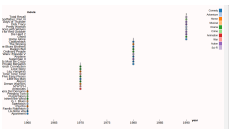
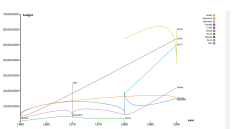
When generating visualizations, arbitrary selection of data fields and subsequent naive choice of visual encoding inevitably lead to a combinatorial explosion. We introduce a novel approach –called VizRec, that relies on perceptual guidelines to reduce the number of recommendations. Considering just visual encoding rules reduces the combinatorial problem, but still leads to a large set of possibilities, valid in terms of representing the data visually, but without considering which type serves user’s needs or preferences [MHT<sup>+</sup>14]. For example, the bar chart in Table 5.1 with the item-id 541 shows the yearly distribution of each movie’s budget (generated for bar chart pattern (2), see Section 3.1.3) without displaying to which movie the budget belongs to. To promote just the relevant views, appropriate filtering and



**Figure 5.1:** Schematic representation of the VizRec recommendation pipeline: The stages (a), (b) and (c) illustrate the preprocessing unit. The stage (d) illustrates the visual mapping process between the elements and the visual patterns, whereby the defined mapping combinations are shown in stage (e). For the personalized visualization recommendation VizRec uses either the user ratings, user and item profiles or a combined version of both (f). Finally, the recommendations will be presented to the user in a top-n manner (g).

recommendation strategies are needed [WMA<sup>+</sup>16]. Together, they should provide only those visualizations that a user would select as part of her analysis workflow. Therefore, we investigate which information lets us anticipate the choice of visualization for data analysis and how to represent such information and use it for recommendation.

The first approach we investigate is collaborative filtering (CF) [SFHS07], which relies on explicit feedback provided by the user in form of ratings. However, ratings alone do not tell much about the content of the data that a visualization represents. To take this aspect into account, we investigate content-based filtering [LdGS11]. Content-based filtering requires metadata information, e.g., in the form of keywords, comments or tags provided by the user. VizRec uses tags, as they have been shown to be useful in many recommender or information retrieval scenarios [LTP<sup>+</sup>15]. Finally, VizRec includes a hybrid recommendation approach, that combines information on what a visualization is about (tags) and how good it is (ratings). This recommendation pipeline including different recommendation strategies defines an additional unit, *user preference filtering* unit, in VizRec’s architecture (see Figure 5.1). The following sections detail the recommendation strategies included in VizRec’s recommendation pipeline.

Item-id	User-id	Rating	Item
254	1	4.0	
	6	4.5	
	10	5.5	
960	1	6.5	
	6	5.5	
	10	4.0	
541	1	2.5	
	6	3.0	
	10	3.5	
721	1	1.0	
	6	2.5	
	10	2.0	
360	6	5.5	
	10	6.5	
...	...	...	...

**Table 5.1:** Input data for the calculation of the k-nearest neighbors and generating predictions for the active user, including item-id, user-id and ratings.

### 5.1.1 Collaborative Filtering

To filter the mapping combinations  $M$  based on the user preferences (see Figure 5.1 (f)), we employ a simple user-based collaborative filtering (CF) approach utilizing ratings [SK09]. The basic idea behind CF is to find a user with similar preferences to the active user, who has rated the item  $x$  that the active user has not seen yet. Hence, the average ratings of the similar users are applied to predict if the active user will prefer the item  $x$ . In a nutshell, the algorithm needs to identify users similar to

the active user, k-nearest neighbors respectively, who share active user's tastes. To calculate the k-nearest neighbors, we construct a  $m \times n$  matrix  $A$  where each entry  $a_{i,j}$  represents the rating of the  $i$ th user on the  $j$ th item (=a mapping combination that defines a chart). Each rating is a numerical scale, e.g., from 1 to 7. Having constructed the matrix  $A$  we employ the Pearson correlation coefficient to calculate the similarity between the active user  $u$  and the user  $v$  using the Equation 5.1:

$$sim(u, v) = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}} \quad (5.1)$$

where  $I$  is the set of items rated by users  $u$  and  $v$  and  $\bar{r}_u$  is the average rating of the active user  $u$ . Once the k-nearest neighbors are detected, VizRec combines the preferences of the neighbors to generate the predictions or the top-n recommendations for the active user, the set  $R$  respectively, following the Equation 5.2:

$$pred_{cf}(u, i) = \bar{r}_u + \frac{\sum_{v \in N} sim(u, v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in N} sim(u, v)} \quad (5.2)$$

where  $\bar{r}_u$  and  $\bar{r}_v$  are the average ratings of the user  $u$  and  $v$ . To do so, VizRec selects from the set of  $M$  only those mappings which the active user might prefer and presents them to her as recommendations (see Figure 5.1 (g)). The list of recommendations  $R$  for the active user is nothing else but a subset of  $M$ .

The cooperation of the mapping algorithm with CF offers two important benefits: first, the definition of perceptually valid visualizations for the active user's dataset and second, the recommendation of only the valid visualizations which the active user might prefer.

To clarify this, we consider our example about the top-ranked movies again. The visual mapping unit defines for this dataset a list with 18 possible bar chart configurations. However, the system contains also alternative visualizations, such as timeline, line chart which are also appropriate for the current dataset.<sup>1</sup> When including their mapping combinations, 6 for timeline and 4 for line chat respectively, the total sum of available visualizations would be 28. Since the preferences of the active user  $u$  are known by the system, it can reduce the list on those to which user's preferences matches the best. To do so, the system first, performs a k-nearest neighborhood search by employing the Pearson correlation measure to detect those users who are the most similar to the active user  $u$ . Table 5.1 shows an excerpt of our example rating database containing ratings for items (visualizations) which has been seen by the active and/or other users whereby the active user is assigned the user-id 1. Second, using the Equation 5.1 for Pearson correlation and the Table 5.1 with item-ids, user-ids and their ratings, the system reveals following similarity values for the active user  $u$ :

$$sim(1, 6) = \frac{(4.0 - 3.5)(4.5 - 3.875) + \dots + (1.0 - 3.5)(2.5 - 3.875)}{\sqrt{(4.0 - 3.5)^2 + \dots + (1.0 - 3.5)^2} \sqrt{(4.5 - 3.875)^2 + \dots + (2.5 - 3.875)^2}} = 0.9806$$

<sup>1</sup>Geo charts were not generated, since the dataset does not contain spatial information

$$\text{sim}(1, 10) = \frac{(4.0 - 3.5)(5.5 - 3.75) + \dots + (1.0 - 3.5)(2.0 - 3.75)}{\sqrt{(4.0 - 3.5)^2 + \dots + (1.0 - 3.5)^2} \sqrt{(5.5 - 3.75)^2 + \dots + (2.0 - 3.75)^2}} = 0.6154$$

Having detected the similarity values for *user6* and *user10* the system tries to predict whether or not the active user *u* might prefer the visualization with the id 360 she has not seen before:

$$\text{pred}_{cf}(1, 360) = 3.5 + \frac{0.9806(5.5 - 3.875) + 0.6154(6.5 - 3.75)}{0.9806 + 0.6154} = 5.559$$

Considering the prediction value for the item 360, we can assume that this item might be one of the ten items being recommended to the active user. To finally define the list of top-10 recommendations, we apply this approach to every user and item the active user has not seen before. Using the similarity values, we rank each item *i* of the *k* most similar users to the active user and present her only the visualizations with the highest ranking.

## 5.1.2 Content-based Filtering

The CF recommendation strategy (CF-RS) needs user's interests beforehand, which should be in common to at least a few other users. When CF-RS cannot find similar users, that is for instance the case when the user or the item is new to the system a content-based recommender (CB-RS) is a suitable alternative. The simplified workflow of the VizRec CB-RS is illustrated in the main Figure 5.1 (f). In a nutshell, the VizRec CB-RS generates recommendations by analyzing the relevant content, concretely, the information we know about the active user and the information we extracted from the items. Following the basic principles of CB-RS, the recommendations are produced based on the content similarity, in our case between the interests of the active user i.e., her profile and the content of the candidate items (visualizations).

### 5.1.2.1 User and Item Profiles

Each visualization generated in VizRec is described with a mapping from metadata to visual components. The metadata elements provide basic information about the content of the visualization they describe. Thus, we build the item profiles based on the current set of mapping combinations (visualizations) the active user observes from the Visual Mapping stage. For instance, when user's dataset is about the *movies* and the mapping combination

$\{x - \text{axis} : \text{movie name}, y - \text{axis} : \text{budget}, \text{color} : \text{genre}\}$  (cf. Figure 3.7 (c)) is one of the candidates, VizRec uses the related metadata elements, *movie name*, *budget* and *genre* to profile this particular mapping combination. Yet, a user annotates items with tags which describe the content of the items and thus serve also as appropriate inputs for their profiles [BVdB09, LTBH15]. To take this into account,



VizRec extends the item profiles with tags supplied by users in the past. To relate the item profiles with the content in the repository, we build an item profile by aggregating the tags of all users per item (visualization). The benefits of the tag aggregation among all users is (i) to obtain more valuable information about individual items than focusing just on information from a single source and (ii) to consider the preferences of the community and not just those of a single user which in turn increases the likelihood that we select items which are relevant for the active user (cf., [BVdB09]). In summary, we build a profile of an item (visualization) based on (i) the metadata they contain, including (if available) (ii) the tags the user supplied to this item in the past. Similar to the item profiles, the user profiles are built upon the tags, the user applied. Usually, user's tags reflect her interest and needs. Thus, we profile each user with her tags which are assumed to describe her interest in a topic or/and an item.

One important concern regarding tags for the user and item profiles is a normalization process which is executed before storing tag information in repository. This process involves, (i) removing of commoner morphological and inflectional endings from English words (e.g., movies  $\rightarrow$  movie, comedies&comedy  $\rightarrow$  comedi) using the Porter stemmer algorithm [KG13], (ii) removing of stop words (standard tokenizer) and punctuations (keyword tokenizer) and finally (iii) the lowercase filtering. This step helps to avoid that the words represented in various language forms are interpreted differently [LdGS11].

### 5.1.2.2 Similarity Estimation and Item Ranking

To determine the correlation between visualizations and users, we transform the content of the user profiles and item profiles into the Vector Space Model (VSM) with the TF-IDF (Term Frequency-Inverse Document Frequency) weighting schema. As mentioned in section 2.1.3.2, VSM is a common technique to vectorize the content and in this way to enable their analysis, such as classification and clustering for example. In our case, VSM consists of user profile (tags) and item profile (mapping combinations), both represented in form of vectors. Concretely, using this scheme, each mapping combination (e.g., *movie name, budget, genre*) is defined as an n-dimensional vector, where each dimension corresponds to a term or more precisely, to the TF-IDF weight of that particular term. To clarify this, let  $M = \{m_1, m_2, m_3, \dots, m_N\}$  be a set of mapping combinations and  $T = \{t_1, t_2, t_3, \dots, t_n\}$  a set of terms in  $M$ . Each mapping combination  $m_i$  is represented as a vector in a n-dimensional vector space, i.e.,  $m_i = w_{1,i}, w_{2,i}, w_{3,i}, \dots, w_{n,i}$ , where  $w_{k,i}$  denotes the weight for the term  $t_k$  in a mapping combination  $m_i$ , i.e.:

$$w_{k,i} = tf - idf_{t_k, m_i} = tf_{t_k, m_i} \times idf_t = tf_{t_k, m_i} \times \left[ \log_e \left( \frac{N}{df_t + 1} \right) + 1 \right] \quad (5.3)$$

where the former factor of the product is an occurrence frequency of the term  $t_k$  within a mapping combination  $m_i$  and the later indicates the distribution of the term

among the both profiles (i.e., so that particular and commonly occurring terms can be discriminated from each other). We apply the same weighting scheme to define the user profile. Having defined the profiles, it is now possible to estimate their similarity. To do so, we use the weighting information in vectors and apply the *cosine similarity* measure [LdGS11], defined as follows:

$$\text{sim}(m_i, m_j) = \frac{\sum_k \mathbf{w}_{k,i} \mathbf{w}_{k,j}}{\sqrt{\sum_k (\mathbf{w}_{k,i})^2} \sqrt{\sum_k (\mathbf{w}_{k,j})^2}} \quad (5.4)$$

where  $m_j$  denotes the tag collection of the current user. The result of this measure is a cosine value of the angle between two vectors, in our case between the mapping combination and the tag collection. The retrieved values are then used as scores to rank the relevant visualizations.




Returning to the example about top-ranked movies filmed in certain period of time. When now including the shooting location (country) and the population of each country into the results, the mapping algorithm would produce a total of 55 visualizations for 4 types of visualizations for the active user– incl. geo chart. Yet, to ascertain which of these visualizations the current user would prefer the most, we define user’s profile by aggregating her tags. An excerpt of the user profile is shown in top part of the Table 5.2. Subsequently, the item profiles are defined by extracting relevant terms (metadata elements) from the individual visualizations. The summarized terms represent here the actual content in each particular visualization, for instance, *genre*, *movie* and *gross* dimensions of the dataset are displayed in the bar chart (the second row). Note, that this profile information can be augmented by additional tags, if available. An excerpt of the item profile generated from metadata elements of each visualization is shown in the bottom part of Table 5.2

The TF-IDF vectors are shown next to the terms in Table 5.2. For example, the tag *genre* in the first mapping combination from the table has a TF-IDF weight of 1.749. The term occurs only once in this mapping, i.e.,  $tf_{genre, mapping38} = 1$  and in 25 other mappings out of 55 overall mappings, i.e.,  $df_{genre} = 25$  and  $N = 55$  respectively, so that  $tf - idf_{genre, mapping38} = 1 \times (\log_e(\frac{55}{25+1}) + 1) = 1.749$ . Note, that here stemmed words are used instead of original ones (i.e., *movie*  $\rightarrow$  *movi*).

In the final step, scores are assigned to individual mapping combinations by comparing vectors of these mappings with vectors from the user profile using Equation 5.4. Concretely in this example, the best matching could be found between mappings 25 (user profile) and 46 (item profile). In a nutshell, their cosine value would be estimated as follows:

$$\text{sim}\left(\begin{bmatrix} 1.749 \\ 1.541 \\ \dots \\ 1.510 \end{bmatrix}, \begin{bmatrix} 5.791 \\ 2.670 \\ \dots \\ 0 \end{bmatrix}\right) = \frac{1.749 \cdot 5.791 + 1.541 \cdot 2.670 + \dots + 1.510 \cdot 0}{\sqrt{1.749^2 + \dots + 1.510^2} \cdot \sqrt{5.791^2 + \dots + 0^2}} = 0.345$$

Hence, the mapping combination 46 with the tags *genre*, *year*, *gross* and *barchart*

User profile		
ID	Terms (frequency)	TF-IDF Vector (weight)
25	movie(5), revenue(2), genre(12), gross(8), collections(2), films(2), most(1), successful(1), film(2), top(1), grosser(1), movies(5), years(1), earnings(1), year(2), money(1), decades(2), profit(1), box(1), office(1), profits(1), decade(1), genres(2)	movi:(4.130), revenue:(3.974), genr:(5.791), gross:(5.420), col- lect:(4.548), film:(3.498), most:(3.216), success:(4.602), top:(3.061), grosser:(4.314), year:(2.670), earn:(2.928), monei:(3.503), decad:(4.298), profit:(4.232), box:(2.810), offic:(2.810)
Item profile		
ID	Terms (frequency)	TF-IDF Vector (weight)
..	...	...
 38	genre(1), movie(1), gross(1), ... , barchart(1)	genr:(1.749), movi:(0.749), gross:(2.174), ... , barchart:(1.510)
 46	genre(1), year(1), gross(1), ... , barchart(1)	genr:(1.749), year:(1.541), gross:(2.174), ... , barchart:(1.510)
 47	genre(1), year(1), budget(1), ... , barchart(1)	genr:(1.749), year:(1.541), bud- get:(2.174), ... , barchart:(1.510)
..	...	...

**Table 5.2:** An excerpt of the *movies* dataset with generated user and item profiles.

might be in this case one of the top-n preferred visualizations for the active user. Finally, having obtained the similarity values for each mapping combination we define the top-n visual recommendations, following the Equation:

$$pred_{cb}(m_i, m_j) = \sum_{m_i, m_j \in M} sim(m_i, m_j) \quad (5.5)$$

### 5.1.3 Hybrid Filtering

The two pieces of information used for recommendation separately describe: what a visualization is about (CB-RS) and how good it is (CF-RS). A combination of these

pieces of information in a single recommendation strategy would arguably supply more meaningful recommendations in varying situations (e.g., when the user or item is new, when user’s interest changes). In general, there exist different methods for a hybrid design [Bur02, JZFF10] (see Section 2.1.3.3).

Item	$rec_1$ score	$rec_1$ rank	$rec_2$ score	$rec_2$ rank	hybrid score	hybrid rank
Item1	0.3	2	0.7	1	0.5	1
Item2	0.5	1	0.3	3	0.4	2
Item3	0.2	3	0.5	2	0.35	3
Item4	0		0.1	4	0.05	

**Table 5.3:** Recommendations for the weighted hybrid recommender.

For the current investigation we have chosen a weighted hybridization design as a first approach in VizRec to utilize the strength of both collaborative filtering and content-based recommender techniques in the straightforward way. Concretely, a weighted hybrid recommender defines the score of a recommended item from the results of all integrated recommender techniques by computing a weighted sum of their scores. When linearly combining the scores, the collaborative and content-based recommenders obtain equal weights. Thus, we use the uniform weighting scheme with  $w_1 = w_2 = 0.5$  for our hybrid recommender and define a new ranking for the recommended items by combining their (normalized) scores from collaborative and content-based recommender following the Equation 5.6.

$$pred_{hyb}(u, i) = \sum_{j=1}^n w_j rec_j(u, i) \quad (5.6)$$

To clarify this process, we consider the table 5.3 containing scores and rankings for five exemplary items. According to this table,  $rec_1$  (CF-RS) produced for *Item1* rank 2 and  $rec_2$  (CB-RS) the rank 1 considering the scores 0.3 and 0.7. When linearly combining those scores following the Equation 5.6, we obtain for the *Item1*:

$$pred_{hyb}(u, Item1) = 0.5 \times 0.3 + 0.5 \times 0.7 = 0.5 \quad (5.7)$$

as the final score. Having computed the hybridized scores for the remaining items, the *Item1* will be finally ranked highest following the *Item2* and *Item3*.

## 5.2 Use-Case: Visual Recommender for Learning in Communities

In today’s digital age we can observe an increasing number of social platforms. Many people use them to build social relations with people who share their tastes, interests and expertise. Such applications also inspired educational designers, researchers and

teachers, especially when they focus on bringing groups of people - communities - to work and learn together. This process is also known as collaborative learning.

Collaborative learning platforms are increasingly being used in education. The primary aim of such platforms is helping participants to share their ideas and information with each other and work together on a task. As a result, the participant's learning process enhances the produced results [Sta04]. However, the level of participation between the group members varies. Where some of the participants highly contribute in discussions and sharing knowledge and resources, some of them put less effort into collaboration. One solution to increase the collaboration might be providing an overview about the productivity of each participant, giving some sort of feedback, using visualizations. For example, when showing the number of messages, answers and shared resources, the participant can easily monitor her collaboration but also follow the traffic surrounding the group she is collaborating in. As a result, the user might take more responsibilities for the own success as well as for the whole group. The same participant can use the visualizations to articulate thoughts about the amount of collaboration within the group. For example, having analyzed the visualizations which show the distribution of the collaboration of each group member, the user may feel that someone is a free rider which could lead to discussing this within the group. As a consequence, the following discussion could raise this participant's awareness of the entire group process. Furthermore, considering the overview of the group process and activities, the teacher or the administrator of the platform can observe if provided strategies for the collaboration are working as expected, if the quality of comments, shared documents are increasing or decreasing over the time and why students put less (or more) effort into collaboration.

There are several factors affecting user's collaboration: knowledge and education level, the quality of the knowledge and resources being shared, her interest in the topic, the politeness in the communication with the peers, just to name a few [JEKJ07]. Typically, the activities of a novice are less at the beginning. It is restricted to e.g., download the resources, read the comments or just follow the activities of the peers. However, when the novice already consumed some resources, her participation should become more active, including e.g., asking questions, writings comments or sharing new resources. If this is the case, we hypothesize that the user achieved some progress in the learning process and could increase knowledge or even acquire new skills. To motivate the user further in collaboration we can provide a visual overview about learners progress. This self-assessment tool might be a timeline showing the increased number of user's learning activities from the beginning by now. In contrast, when user's collaboration remains low, the topic and the provided resources are maybe too low or high level. This can be assessed by considering the overall motivational outcomes in the group. By (visually) considering the distribution of the ratings (given to the resources incl. comments) indicating the level of satisfaction we can draw conclusion about why the collaboration of new users are less.

In summary, visualizations facilitate discovering, analyzing and understanding

communities. They help us to analyze the characteristics of the communities, to understand how learning happens there and if and how this process is supported. Furthermore, visualizing learners' activities supports self-awareness and reflection. It enables learners to discover the learning activities of the peers, to compare own progress with those of the peers respectively. Consequently, this may increase the awareness and lead learners to invest more effort into the collaboration.

There exist several systems visualizing learner activities in learning platforms, categorized into those which reflect actions and those which monitor the interaction state [GVKD10]. The systems from the first category provide supports which make private areas of the learner's workspace respectively. The systems from the second category either aggregate user's interactions, such as the number of messages the user applied, into a set of high-level indicators and show them to the user or compare the current state of the interaction to a model of ideal interaction but do not show this to the user. Each of the above-mentioned systems has different objectives: increasing awareness, supporting self-reflection, improving collaboration among learners and discovering the learning activities of the peers.

We propose VizRec which covers mainly the first and partly the second objective to support learning analytics: visualizing the learning activities and providing statistics about the activities. Furthermore, our system provide user the controllability about which aspects of the data has to be considered and also about which statistics have to be defined.

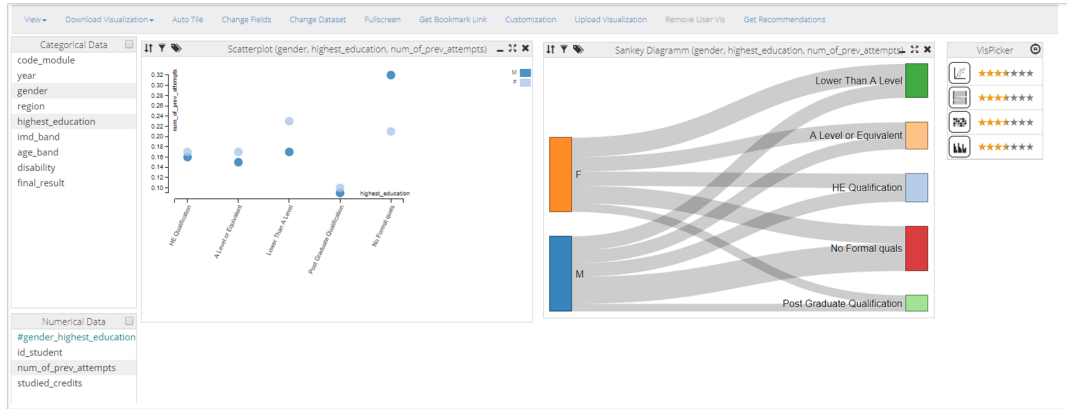
In contrast to the existing systems [VGD<sup>+</sup>14] VizRec is context sensitive (data, persons, tasks) and supports personalization. The system is able to automatically define a set of appropriate visualization considering the features of the log data (containing user's learning activities) and further to filter a subset based on target visual preferences, interest and task. Note, a target user in our context is either (i) a learner who considers her personal data or those of the peers or (ii) a teacher/administrator who considers the data of a specific user or of the entire group.

Currently, VizRec is used in EU-funded AFEL (Analytics for Everyday Learning) project.<sup>2</sup> "The goal of AFEL is to develop methods and tools for exploiting learning analytics on learning activities in online social environments."<sup>3</sup> VizRec's role in this project however is to serve as a personalized learning dashboard providing standard charts (e.g. bar/pie charts, scatter plots, box plots, line charts etc.) for monitoring and assessing multi-dimensional statistical information collected during learning activity (e.g. learning progress, view frequencies etc.). The framework is designed to provide overview information on learner, community and learning resource information using the most suitable chart(s) considering characteristics of the data being visualized and visual preferences and needs of the learners/teachers. The Figure 5.2 illustrates a usage scenario for the personalized learning dashboard. In this scenario, a teacher explores the number of previous attempts of females and males who take her class with regard to their education level. To do so, she

---

<sup>2</sup><https://vizrectest.know-center.tugraz.at/>

<sup>3</sup><http://afel-project.eu/>



**Figure 5.2:** Personalized learning dashboard: Teacher explores the number of previous attempts of females and males who take her class with regard to their education level. The teacher first selects the data attributes of interest (see left) the system then recommends the personalized visualizations (see right). One selected, the personalized visualizations are created automatically (see center).

first selects the data attributes she is interested in (i.e., gender, highest\_education, num\_of\_previous\_attempts) (see Figure 5.2 left), VizRec then recommends the visualizations that are closer to what she prefers and needs (see Figure 5.2 right). One selected, the personalized visualizations are created automatically (see Figure 5.2 center). Looking at the created visualizations, the teacher can easily see that, for instance, the females with an educational level lower than A have had more previous attempts compared to the male colleagues.

## 5.3 Evaluation

In this section we investigate the performance of the individual recommendation strategies. To do so, we designed a study on a crowd-sourced platform to elicit preferences for a fix number of visualizations associated with three different datasets (Movies, EU, Books). Note, the details about the study and datasets can be taken from Chapter 4. We then performed several offline evaluations to study the individual recommenders. This section describes in detail the method and metrics used and the results of the studies.

### 5.3.1 Evaluation Protocol

We performed offline experiments to estimate the quality of personal preferences for visualization recommendations. To this end, we used the preferences collected from the Amazon study as input data to train our recommenders. For the CF-RS, we maintain a list of items (visualizations), each having the information about user and

provided rating. Similarly, CB-RS uses the tags per item. Finally, for the hybrid approach we combine the results of both recommender techniques. Following the method described in [TKL15], we split the preference model into the two distinct sets, one for training the recommender (training-set) and another one for testing (test-set). The test-set acts here as a reference value that, in an ideal case, has to be fully predicted for the given training-set. From each of the datasets in the preference model, we randomly selected a certain percent (more details are given in Section 5.3.2) of user-rated or user-tagged mapping combinations (visualizations) and entered them into the training-set performing five-fold cross validation. The recommendations produced out of the training-set are further used to evaluate the performance of VizRec. The performance of VizRec depends generally on how good it predicts the test-set. We compared the generated recommendations (prediction-set) and the test-set by applying a variety of well-known evaluation metrics in information retrieval [HKTR04]: Recall ( $R$ ), Precision ( $P$ ), F-Measure ( $F$ ), Mean Average Precision ( $MAP$ ) and the Normalized Discounted Cumulative Gain ( $nDCG$ ). The first three metrics basically express the quantity of relevant recommended results, whereas  $MAP$  and  $nDCG$  quantify the concrete ordering of the results (i.e., penalizing results which are not on the top but are relevant for the user). Concretely, the metrics are defined as follows:

*Recall* ( $R@k$ ) is calculated as the number of correctly recommended visualizations divided by the number of relevant visualizations, where  $r_u^k$  denotes the top  $k$  recommended visualizations and  $R_u$  the list of relevant visualizations of a user  $u$  in the set of all users  $U$ . Recall is given by [Rij74]:

$$R@k = \frac{1}{|U|} \sum_{u \in U} \left( \frac{|r_u^k \cap R_u|}{|R_u|} \right) \quad (5.8)$$

*Precision* ( $P@k$ ) is calculated as the number of correctly recommended visualizations divided by the number of recommended visualizations  $k$ . Precision is defined as [Rij74]:

$$P@k = \frac{1}{|U|} \sum_{u \in U} \left( \frac{|r_u^k \cap R_u|}{k} \right) \quad (5.9)$$

*F1-score* ( $F1$ ) combines precision and recall into one score [Rij74]:

$$F1@k = 2 \cdot \frac{P@k \cdot R@k}{P@k + R@k} \quad (5.10)$$

*Mean average precision* ( $MAP$ ) is an extension of the precision metric that additionally looks at the ranking of recommended visualizations. MAP is described in the subsequent equation, where  $B_j$  is 1 if the recommended visualization at position  $j$  is among the relevant visualizations and 0 otherwise [RKAES13]:

$$MAP@k = \frac{1}{|U|} \sum_{u \in U} \frac{1}{k} \sum_{j=1}^k B_j \cdot P@j \quad (5.11)$$



*Normalized Discounted Cumulative Gain* ( $nDCG@k$ ) is a ranking-dependent metric that not only measures how many visualizations can be correctly predicted but also takes the position of the visualizations in the recommended list with length  $k$  into account. The  $nDCG$  metric is based on the *Discounted Cumulative Gain* ( $DCG@k$ ) which is given by [PS13]:

$$DCG@k = \sum_{i=1}^k \left( \frac{2^{B(i)} - 1}{\log_2(1 + i)} \right) \quad (5.12)$$

where  $B(i)$  is a function that returns 1 if the recommended product at position  $i$  in the recommended list is relevant.  $nDCG@k$  is calculated as  $DCG@k$  divided by the ideal DCG value  $iDCG@k$  which is the highest possible DCG value that can be achieved if all the relevant visualizations would be recommended in the correct order. Taken together, it is given by the following equation [PS13]:

$$nDCG@k = \frac{1}{|U|} \sum_{u \in U} \left( \frac{DCG@k}{iDCG@k} \right) \quad (5.13)$$

### 5.3.2 Recommendation Quality

In this section, we summarize the results of the offline evaluations. As defined in the protocol, first, we show how VizRec performs with regard to user preferences collected in Amazon Turk experiment. We analyze here how the recommender performs using individual rating dimensions compared to the performance with the aggregated ratings (overall score). Second, we show what kind of user feedback, rating values or annotations via tags, would be more adequate for recommending visualizations. In addition, we compare the performance of collaborative filtering based recommender and content-based recommender techniques with their hybridized version (hybrid recommender).

#### 5.3.2.1 Using Each Rating Dimension Separately

This part of the experiment was intended to compare VizRec’s recommender performance when using a single rating vs. using a multidimensional scale. To do so, we estimated the quality metric values Recall ( $R$ ), Precision ( $P$ ), F-Measure ( $F$ ), Mean Average Precision ( $MAP$ ) and the Normalized Discounted Cumulative Gain ( $nDCG$ ), for each of the individual ratings and for the overall score. Table 5.4 summarizes the results.

The results show that recommendations generated with O.a (overall score) are more accurate than those obtained with either of the nine dimensions separately. For instance, when comparing by dataset *Movies* the recommendation accuracy ( $F@3$ ) for dimension UN with the value for the overall score, the dependent  $t$ -test reveals that VizRec’s CF performs, on average, significantly better for the overall rating

( $M = 0.1137, SE = 0.0077$ ) than for the dimension UN ( $M = 0.0632, SE = 0.0036$ ),  $t(35) = 2.5204, p < 0.01, r = 0.400$ . Subsequently,  $MAP@3$  ascertains that when using the overall rating ( $M = 0.1011, SE = 0.0063$ ) VizRec can sort individual recommendations according to their relevance to the user significantly better than e.g, using the dimension UN ( $M = 0.0345, SE = 0.0038$ ),  $t(35) = 2.6759, p < 0.01, r = 0.41$ . Note, the effect size estimate ( $r$ ) indicates that the difference in performance is a large and therefore a substantive, effect (just below 0.5) – all effects are reported at a 0.05 level of significance.

The results support our assumption that considering different aspects to rate visualizations improves recommendation quality. This has a root in the fact that individual dimensions are potential source of errors, as user may understand and interpret them in different ways. In addition, when providing rating values, there is often a need for a reference value, based on which such absolute ratings can be made (e.g., when just taking the subjective judgment on "useful" for the first time). On the contrary, different aspects may provide such a reference value, as user get insight on what else may be required that eventually stays in relation with other dimensions (e.g., easy to understand and confusing). Furthermore, it is more likely that additional dimensions will compensate for mistakes on individual dimensions, like being unable to evaluate it objectively.

Another finding here is, that there is no pattern across the nine dimensions implying a dependence of the recommendation accuracy on negative (boring, cluttered, etc.) or on positive (effective, exciting, etc.) dimensions. For instance, for the *Movies* dataset the  $F$  – *Measure* for the positive dimension *easy to understand* is  $F@3_{UN} = 0.0632$  whereby for its opposite dimension *confusing*  $F@3_{co} = 0.0891$ . A dependent  $t$  – *test* reveals that the recommendation accuracy for *confusing* ( $-co$ ) ( $M = 0.0891, SE = 0.0067$ ) is, on average, not significantly higher than for *easy to understand* (*UN*) ( $M = 0.0632, SE = 0.0036$ ),  $t(35) = 1.4146, p > 0.01, r = 0.2325$ . The effect size estimate indicates that the difference in recommendation accuracy given by negative dimensions  $-co$  is a small and therefore unsubstantial effect. The same effect is present for the positive dimension *exciting* (*EX*) ( $M = 0.1003, SE = 0.0130$ ) and its opposite dimension *boring* ( $-bo$ ) ( $M = 0.0854, SE = 0.0079$ )  $t(35) = 0.7042, p > 0.01, r = 0.1181$ . These results indicate that no dimension dominates the others and thus has a special impact on the overall rating. In summary, negative ratings are as valuable input as the positive ratings [SFHS07] but as many recommender systems, VizRec performs better using both positive and negative ratings.

### 5.3.2.2 Using Overall Scores

To measure the improvements in terms of recommender quality (= accuracy), we compared the VizRec CF with the baseline filtering algorithms Most Popular (MP) and Random (RD). The RD simulates the recommender behavior providing an arbitrary order of visualizations – i.e., it can be compared with having only the first two units in the VizRec pipeline from Figure 5.1. The MP, in contrast, generates the

Dataset	Dimensions	Metric				
		R@3	P@3	F@3	MAP@3	nDCG@3
Movies	-bo	0.0814	0.1425	0.0854	0.0598	0.0924
	-cl	0.0551	0.1389	0.0757	0.0419	0.0761
	-co	0.0868	0.1481	0.0891	0.0548	0.0920
	EF	0.0905	0.1629	0.0972	0.0638	0.1027
	EX	0.0993	0.1592	0.1003	0.0687	0.1071
	OR	0.0692	0.1463	0.0866	0.0584	0.0872
	SA	0.0834	0.1481	0.0895	0.0696	0.1022
	UN	0.0470	0.1185	0.0632	0.0345	0.0642
	US	0.0983	0.1537	0.0970	0.0620	0.1028
	O.a	0.1320	0.1685	0.1137	0.1011	0.1362
EU	-bo	0.2080	0.3473	0.1286	0.1800	0.2488
	-bl	0.2592	0.3649	0.2754	0.2061	0.2833
	-co	0.2259	0.3789	0.2745	0.1785	0.2540
	EF	0.2471	0.3754	0.2783	0.2005	0.2768
	EX	0.2203	0.3684	0.2687	0.1814	0.2555
	OR	0.2107	0.3614	0.2588	0.1764	0.2511
	SA	0.1884	0.3403	0.2392	0.1691	0.2348
	UN	0.2080	0.3614	0.2589	0.1859	0.2551
	US	0.2270	0.3649	0.2640	0.1945	0.2615
	O.a	0.2701	0.3684	0.2801	0.2199	0.2954
Books	-bo	0.5888	0.4259	0.4677	0.4629	0.4949
	-cl	0.6666	0.5155	0.5573	0.5711	0.5980
	-co	0.6066	0.4888	0.5182	0.5333	0.5513
	EF	0.5222	0.3518	0.3955	0.3833	0.4186
	EX	0.5466	0.4906	0.4955	0.1814	0.5110
	OR	0.6133	0.4488	0.4920	0.4544	0.4944
	SA	0.5466	0.3822	0.4266	0.4377	0.4675
	UN	0.6400	0.4844	0.5266	0.5522	0.5753
	US	0.5444	0.4074	0.4422	0.4592	0.4812
	O.a	0.6933	0.4400	0.5626	0.5966	0.6220

**Table 5.4:** Quality metrics values estimated for the three example datasets using VizRec’s CF-RS. The values are calculated first for the ratings taken from one out of the nine dimensions (bo=boring, cl=cluttered, co=confusing, EF=efficient, EX=exciting OR=organized, SA=satisfying, UN=Easy to understand, US=useful) and then for the overall rating score (O.a). Note, we inverted the ratings of the negative dimensions (N) boring, cluttered and confusing using their opposites (P) exciting organized and easy to understand according to the equation  $\frac{(P-N)+7}{2}$ . For this test we used a 5-fold cross validation whereby each iteration used 80% of user’s data as training set and 20% as test set.

results sorted according to global ratings, in our case accumulated from ratings of individual users. Considering RD and MP, baseline algorithms should unveil whether the recommender systems can in general help with providing useful visualizations and whether the personalized approach improves the quality of the results.

Dataset	Alg.	Metric				
		R@3	P@3	F@3	MAP@3	nDCG@3
Movies	CF	0.1320	0.1685	0.1137	0.1011	0.1362
	MP	0.0488	0.0926	0.0591	0.0163	0.0419
	RD	0.0039	0.0093	0.0055	0.0020	0.0048
EU	CF	0.2701	0.3684	0.2801	0.2199	0.2954
	MP	0.0263	0.0175	0.0211	0.0088	0.0161
	RD	0.0132	0.0175	0.0150	0.0044	0.0103
Books	CF	0.6933	0.4400	0.5626	0.5966	0.6220
	MP	0.1333	0.0444	0.0667	0.0444	0.0667
	RD	0.0667	0.0222	0.0333	0.0333	0.0420

**Table 5.5:** Quality metrics values R@3, P@3, F@3, MAP@3, nDCG@3 estimated for the three datasets using the baseline algorithms MP and RD ( $k = 3$ ). Note, for this test we executed a 5-fold cross validation whereby each iteration used 80% of user’s data as training set and 20% as test set.

Table 5.5 summarizes the results of the evaluation. VizRec CF outperforms both baseline algorithms in all three datasets. The first three quality metrics clearly indicate that the results are more accurate using *VizRec* CF than simply generating arbitrary visualizations (cf.,  $F@3(CF) = 0.1137$  and  $F@3(RD) = 0.0055$  for *Movies*). Concretely, the dependent  $t - test$  reveals that, on average, the performance of the CF ( $M = 0.1137, SE = 0.0077$ ) is significantly higher than the baseline algorithm RD ( $M = 0.0055, SE = 0.0149$ ),  $t(35) = 3.0375, p < 0.01, r = 0.4567$ . The effect size estimate ( $r$ ) indicates that the difference in performance is a large and therefore a substantive, effect. Additionally,  $MAP@3$  and  $nDCG@3$  reveal that *VizRec* CF is significantly better at sorting individual visualizations according to their relevance to the user. For example, the results for  $MAP@3$  show that CF ( $M = 0.1011, SE = 0.0063$ ) significantly outperforms RD ( $M = 0.0020, SE = 0.0053$ ),  $t(35) = 3.9771, p < 0.01, r = 0.0558$ . The effect size estimate indicates also a large difference in performance and therefore a substantive effect – all effects are reported at a 0.05 level of significance.

Note, that the difference between individual metrics among the datasets is to a large extent influenced by the considerable difference in size of the three datasets (e.g., Books has only 7 different visualizations –  $F@3(CF) = 0.5626$ , whereas *Movies* has 55 –  $F@3(CF) = 0.1137$ , see Figure 4.4).

Another interesting finding is that the recommender strategy based on global ratings (MP) generated less accurate results than *VizRec* CF, both with regard to providing relevant visualizations and their ranking order. This supports our main assumption that in terms of the wide variability in user preference ratings, the personalized approach performs better recommendations.

### 5.3.2.3 Collaborative Filtering vs. Content-Based Recommendations

For the sake of comparing both VizRec recommenders, we extended the comparison using the same quality metrics to rating-based (CF) and tag-based (CB) recommender techniques. The estimation was performed in five runs using random splitting of training/test data, as described in the procedure of the experiment. Each iteration uses 80%, 60%, 40% and 20% of user's data as training and the rest as test set. Table 5.6 summarizes the results Figure 5.3 illustrates them. The first observation reveals that both approaches show relatively comparable performance in all three datasets, particularly for smaller training data (i.e., cases with 20% and 40% of the users' data). For the majority of the quality measures in these two cases the CB outperforms the rating-based (CF) recommender technique, both in providing expected recommendations and in sorting. However, in few exceptional cases, mostly for 40% of the users' data, CF shows slightly better results (cf. Figure 5.3 (b), (d), (e) and (f)).

The results also reveal that the relative improvement in recommendation performance depends to some extent on characteristics of the dataset. For example, for *Books* the F-score measure for CB is about 5 times better than CF, while for *Movies* and *EU* datasets the relative improvement in performance lies at about 8% and 9% respectively. Concretely, when considering the recommendation quality ( $F@3$ ) for the dataset *Movies* with 20% of users' data, the dependent  $t$ -test reveals that, CB ( $M = 0.0399, SE = 0.0018$ ) does not perform significantly better than CF ( $M = 0.0368, SE = 0.0069$ ),  $t(35) = 0.5879, p > 0.01, r = 0.0988$ . For the dataset *EU* the test reveals similar results – CB ( $M = 0.1294, SE = 0.0082$ ) does not perform significantly better than CF ( $M = 0.1188, SE = 0.0061$ ),  $t(18) = 0.5800, p > 0.01, r = 0.1354$ . On the other hand for the the dataset *Books* we can observe that the performance of CB ( $M = 0.1622, SE = 0.0091$ ) is significantly better than CF ( $M = 0.0253, SE = 0.0125$ ),  $t(14) = 2.8789, p < 0.01, r = 0.5965$ . The effect size estimate ( $r$ ) indicates that the difference in performance is a large and therefore a substantive, effect.

Conversely, the last case of the experiment—with the 80% of the users' data—, the dependent  $t$ -test for *Movies* reveals that the recommendation quality ( $F@3$ ) with CF ( $M = 0.1137, SE = 0.0077$ ) is, on average, significantly higher than with CB ( $M = 0.0367, SE = 0.0033$ ),  $t(35) = 3.1604, p < 0.01, r = 0.4711$ . The effect size estimate ( $r$ ) indicates that the difference in performance is a large and therefore a substantive, effect. For this training-set configuration, the user has more data (i.e., rated or tagged items) and specific preferences are of more importance than in previous cases with 20% and 40% of the training data. Since the CF looks at other user to find recommendations, more specific results could be observed compared to CB.

In a nutshell, an important finding in this study was that both algorithms behave differently in response to size of user preferences/profiles. With small user profiles/preferences, the tag-based recommender performs better recommendation

		Metric									
Dataset	Alg.	R@3		P@3		F@3		MAP@3		nDCG@3	
		$ts_{20}$	$ts_{40}$	$ts_{20}$	$ts_{40}$	$ts_{20}$	$ts_{40}$	$ts_{20}$	$ts_{40}$	$ts_{20}$	$ts_{40}$
Movies	CB	0.0226	0.0595	0.2111	0.1925	0.0399	0.0693	0.0189	0.0399	0.0514	0.0782
	CF	0.0207	0.0289	0.2018	0.1944	0.0368	0.0482	0.0175	0.0241	0.0467	0.0571
	Hybrid	0.0434	0.1145	0.3407	0.4370	0.0751	0.1658	0.0434	0.1145	0.0916	0.1873
EU	CB	0.0694	0.0694	0.4561	0.3859	0.1294	0.1617	0.0663	0.0983	0.1472	0.1832
	CF	0.0687	0.1012	0.4491	0.4491	0.1188	0.1584	0.0658	0.0906	0.1382	0.1746
	Hybrid	0.0821	0.2043	0.4842	0.6842	0.1372	0.2961	0.0714	0.2043	0.1550	0.3113
Books	CB	0.1226	0.2371	0.2400	0.3066	0.1622	0.2730	0.0822	0.2093	0.1462	0.2750
	CF	0.0177	0.2093	0.0533	0.3600	0.0253	0.2633	0.0177	0.1768	0.0252	0.2536
	Hybrid	0.1408	0.2800	0.4400	0.5733	0.2074	0.3177	0.1408	0.2371	0.2060	0.3154
		$ts_{60}$	$ts_{80}$	$ts_{60}$	$ts_{80}$	$ts_{60}$	$ts_{80}$	$ts_{60}$	$ts_{80}$	$ts_{60}$	$ts_{80}$
Movies	CB	0.0296	0.0310	0.1018	0.0574	0.0400	0.0367	0.0185	0.0186	0.0418	0.0348
	CF	0.0401	0.1320	0.1684	0.1685	0.0600	0.1137	0.0305	0.1011	0.0641	0.1362
	Hybrid	0.1495	0.3988	0.4166	0.3962	0.1998	0.3246	0.1495	0.3988	0.2196	0.4640
EU	CB	0.0694	0.0778	0.2105	0.1087	0.1104	0.0867	0.0643	0.0583	0.1206	0.0965
	CF	0.1322	0.2701	0.4245	0.3684	0.1975	0.2801	0.1167	0.2199	0.2016	0.2954
	Hybrid	0.2903	0.5303	0.6982	0.6561	0.3901	0.5444	0.2903	0.5303	0.3979	0.6073
Books	CB	0.2800	0.2800	0.2400	0.1644	0.2568	0.2160	0.1788	0.2144	0.2612	0.2690
	CF	0.4822	0.6933	0.4488	0.4400	0.4551	0.5626	0.4477	0.5966	0.4623	0.6220
	Hybrid	0.6088	0.8600	0.5511	0.5155	0.5742	0.5680	0.5811	0.8600	0.6048	0.8615

**Table 5.6:** Comparing VizRec hybrid approach with rating-based and tag-based approaches: Quality metric values considering the first three recommendations in the list ( $k = 3$ ). The results are distributed over four sets (from 20% to 80%), each containing different number of items (visualizations) in the training set  $ts$ . Note, for this test we used a 5-fold cross validation.

quality than the rating-based filtering approach, where the results remain stable for almost all quality metrics. With smaller training sets, the tag-based filtering seems to be a method of choice.

### 5.3.2.4 Hybrid Recommendations

To evaluate the performance of our hybrid recommender, we use the same quality metrics and compare the results with those previously estimated for CF and CB. Again, we run the recommender using a 5-fold cross validation. The results are summarized in Table 5.6 and illustrated in Figure 5.3.

Considering recommendation accuracy for all four training/test sets and three datasets, the hybrid recommender outperforms both CF and tag-based CB. For instance, for  $Movies@t_{80}$ , the dependent  $t$ -test between Hybrid and rating-based (CF) recommender reveals, that, on average, the recommendation accuracy for Hybrid ( $M = 0.3246$ ,  $SE = 0.0202$ ) is significantly higher than for CF ( $M = 0.1137$ ,  $SE = 0.0077$ ),  $t(35) = 6.6380$ ,  $p < 0.01$ ,  $r = 0.7465$ . The effect size estimate ( $r$ ) indi-

cates that the difference in performance is a large and therefore a substantive, effect ( $r > 0.5$ ). Furthermore, the dependent  $t$ -test between Hybrid and tag-based (CB) recommender for *Movies@t<sub>80</sub>* delivers similar results. Concretely, the recommendation accuracy for Hybrid ( $M = 0.3246, SE = 0.0202$ ) is significantly higher than for CB ( $M = 0.0367, SE = 0.0033$ ),  $t(35) = 9.6200, p < 0.01, r = 0.8518$ . The effect size estimate ( $r$ ) also indicates a large difference in performance and therefore a substantive, effect ( $r > 0.5$ ).

For the sake of evaluating the performance of Hybrid recommender for more than one dataset, we consider now the recommendation accuracy for the dataset *Eu@t<sub>80</sub>*. The dependent  $t$ -test between Hybrid and rating-based (CF) recommender reveals that, on average, the recommendation accuracy for Hybrid ( $M = 0.5444, SE = 0.0151$ ) is significantly higher than for CF ( $M = 0.2801, SE = 0.0076$ ),  $t(18) = 3.1960, p < 0.01, r = 0.6016$ . The effect size estimate ( $r$ ) for this test indicates that the difference in performance is a large and therefore a substantive, effect ( $r > 0.5$ ). Furthermore, the dependent  $t$ -test between Hybrid and tag-based (CB) recommender for *Eu@t<sub>80</sub>* delivers similar results. Concretely, the recommendation accuracy for Hybrid ( $M = 0.5444, SE = 0.0151$ ) is significantly higher than for CB ( $M = 0.0867, SE = 0.0141$ ),  $t(18) = 8.1385, p < 0.01, r = 0.8867$ . Similar to the previous test, the effect size estimate ( $r$ ) indicates that the difference in performance is a large and therefore a substantive, effect ( $r > 0.5$ ).

Subsequently,  $MAP@3$  and  $nDCG@3$  values @*ts<sub>80</sub>* ascertain that the Hybrid recommender can sort individual recommendations according to their relevance to the user better. Concretely, the dependent  $t$ -test for *Movies@t<sub>80</sub>* between Hybrid and CF (on  $nDCG@3$  values) reveals that Hybrid ( $M = 0.4640, SE = 0.0333$ ) recommender performs significantly better than CF ( $M = 0.1362, SE = 0.0081$ ),  $t(35) = 7.7978, p < 0.01, r = 0.7966$ . Finally, the effect size estimate ( $r$ ) indicates that the difference in performance is a large and therefore a substantive, effect ( $r > 0.5$ ).

An interesting finding here is that the more data exist about the user the better the Hybrid recommender performs (cf., *Movies@ts<sub>80</sub>*:  $F@3_{Hybrid} = 0.3246$ ,  $nDCG@3_{Hybrid} = 0.4640$ , *Movies@ts<sub>20</sub>*:  $F@3_{Hybrid} = 0.0751$ ,  $nDCG@3_{Hybrid} = 0.0916$ ). As already shown in the previous study, this finding does not hold for the CB. On the other side, CF recommender behaves similar to the Hybrid recommender but not in the same volume (cf., *Movies@ts<sub>80</sub>*:  $F@3_{Rate} = 0.1137$ ,  $F@3_{Hybrid} = 0.3246$ ).

In summary, our studies reveal that combining users' tags and ratings improves the quality of recommendations significantly. The ratings and tags are user specific, i.e., ratings present user's general tastes and tags user's topic of interest. When combining both, a recommender system has more detailed information stored in user preferences and can respond more accurately. Concretely, the system can consider larger diversity of item types for defining the prediction, which in turn increases the likelihood that the user will be recommended items which match her preferences the best. When now considering our second evaluation goal – investigate what kind

of feedback is more useful, in terms of recommendation quality, we can finally say that ratings and tags together help to build more accurate recommender system for personalized visualizations.

### 5.3.3 Discussion

Creating and proposing just the relevant visualizations requires appropriate filtering and recommendation strategies. Our investigations build on the premise that preference of a visual representation for data is a personal matter. We set out to investigate which information lets us anticipate the choice of chart, how to represent such information and use it for recommendation. Through a crowd-based experiment we collected empirical evidence supporting the assumption that preferences vary widely for visual representations generated automatically (see Chapter 4, Section 4.2.2). Beyond visual perception guidelines, there are other reasons that lead people to choose a particular representation. It may be habituation, as a user may be comfortable with using a particular representation for data analysis, though this has to be validated in future studies. Our research tries to unveil which aspects have to be present to recommend relevant visualizations. We outlined a rating scale comprising nine dimensions built upon established usability factors. Scores for charts obtained in each of these dimensions were used to train a CF-RS and compared to an aggregated score averaging all nine dimensions. The overall score performed better recommendations compared to our rule based approach and to a most popular method, confirming our assumptions that personalization is important.

These metrics are only based on assessed quality of a chart. However, the choice of representation is also tied to the task or question the user seeks to answer. To represent these aspects, we used tags elicited through the crowd based experiment to train a CB-RS. Comparing both CF and CB leads to the conclusion that tags are good descriptors when there is little knowledge about user preferences. Yet, as tags contribute a great deal of knowledge about the user, we combined both pieces of information in a hybrid recommender approach. The studies revealed that the hybrid approach significantly outperformed both CF and CB in most occasions.

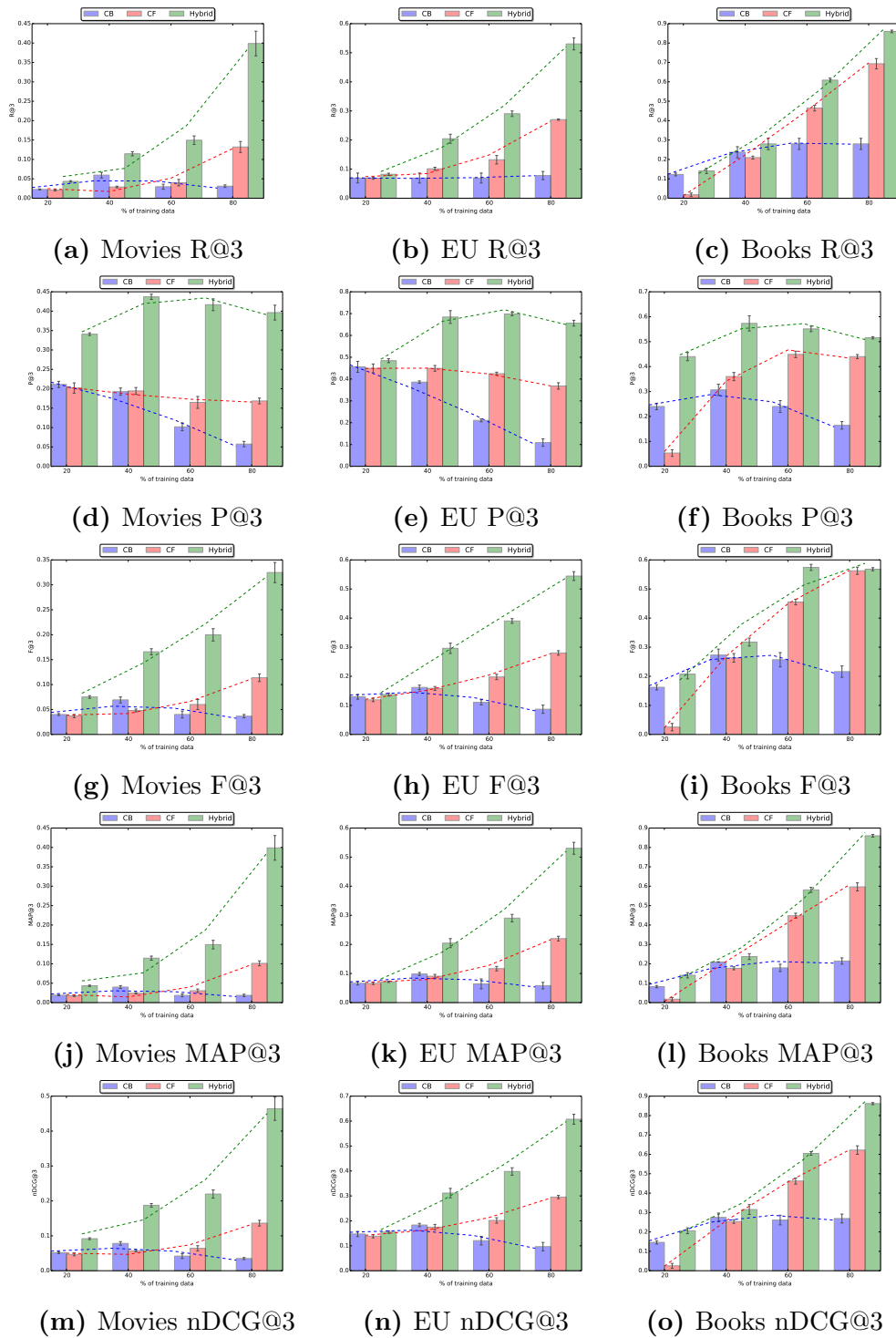
Our research did not concentrate on whether users are willing to provide information (tags/ratings) for visualizations. This is a valid research question for future work (see Chapter 6). Relevant works [VWV<sup>+</sup>07, WSP<sup>+</sup>06] reveal the benefit of annotating visualizations in context of information retrieval. When annotating the user provides her insights and her interpretation on the data being visualized. Hence, they serve as analysis finding records and personal reminder for later data discovering and analysis tasks [EB12]. In the current work, we simply used known information such as the user query and the dataset fields as part of the tag vector describing user's needs and visualization respectively, but clearly better interfaces are needed to make sure that the needed information is there for the recommendation strategy to work.



## 5.4 Summary

The approach presented here is based on the premise that the preferences of a visual representation for a dataset is a personal preference. Empirical evidence collected through a crowd sourced experiment supported the assumption that preferences widely vary for visual representations generated automatically. The studies in this chapter however could reveal that Collaborative Filtering and Content based Filtering approaches can account for such variability in personal preferences and can generate visual recommendations that are rather close to what the user needs and prefers. A major contribution of the presented work is that it is based on the empirical evidence collected following methodical studies involving the general public. Our approach to generating and recommending personalized visualizations, the process of elicitation of users' preferences and the insights described in this thesis are to the best of our knowledge, novel [MVTS15a].

Yet, the effectiveness of personalized recommender systems highly depends on the user and the item profile completeness and accuracy. In next Chapter, we first investigate the power of different kind of user provided input in encoding user's visual preferences (=user profile) and the content of the visualizations (=item profile) and next their practicability in recommending personalized visualizations.



**Figure 5.3:** Performance plots for collaborative filtering (CF), content based (CB) and hybrid approaches for different sizes of training sets (20%-80%) considering the first three recommendations in the result list ( $k = 3$ ).



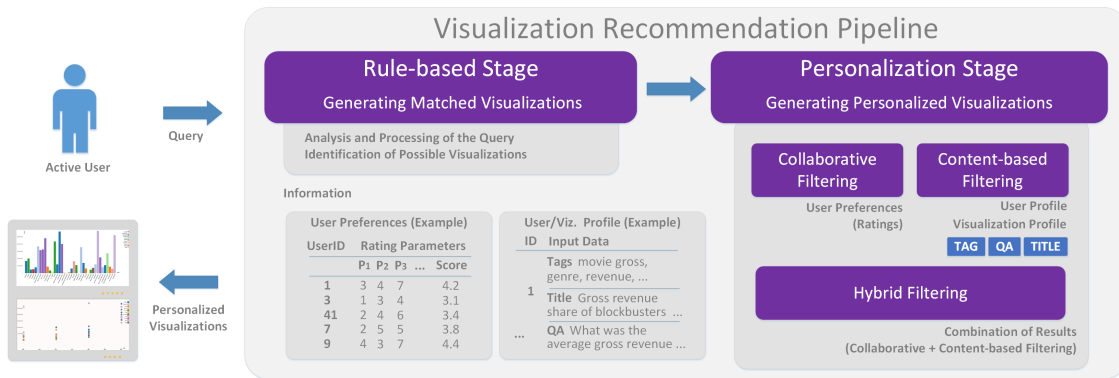
## Chapter 6

# Content Descriptors for Visual Recommender Systems

The effectiveness of personalized recommender systems (RS) highly depends on user and item profile completeness and accuracy [NKA11]. Hence, regardless of the approach, the key factor in personalized recommendations is the decision about how to exploit the relevant information about the user and items and more important, which information better describes the properties of items and the preferences/needs of a user. Collaborative filtering approaches for instance use ratings, while content based approaches build on user-provided input, typically in form of tags and comments. Rating is fast, simple and effective in communicating user preferences also in the context of visualizations [MVT15a], but it does not indicate much about goals or intentions of the user regarding the item. Annotating visualizations with tags brings extra benefits, as a user indicates her insights and interpretation of the data being visualized, i.e., issuing details with keywords pulled from a personal vocabulary [VWV<sup>+</sup>07, WSP<sup>+</sup>06, MVT16]. Hereby, visualizations are organized for later retrieval [MVT17].

There are two caveats in these approaches to personalizing visualizations: 1) people are often reluctant to give a feedback, 2) ratings and tags forego information about the context where the item was used. Unless the benefit is evident, users rarely engage in tagging or rating items. This is true in the context [VWV<sup>+</sup>07, EB12] and can be more acute for visualizations where the user is possibly engaged in a thought process that would be interrupted by rating/tagging. More importantly, a single rating does not tell much about goals or intentions of the user. Whereas tags encode features of the item, it is not evident that users will include their task or intentions when tagging a chart. Our working question is: can we use alternative sources to derive item descriptions suitable for recommendation [MVT17]?

In the context of visualizations, user's provided input (annotations) can take other forms. For instance, it is common for a user to pose a question that is answered with a visualization or to define a title and description for the visualization in form of a caption. We consider these two alternative sources of information (titles,



**Figure 6.1:** Schematic representation of the visualization recommender: The rule-based stage applies visual encoding guidelines to generate a collection of visualizations appropriate for the data. The personalization stage applies the user preferences/profiles (ratings, content terms such as tags and titles) and filters the visualizations according to users' needs and interests. This stage also maintains repositories for the user preferences/profiles.

questions&answers (Q&As)) as potential descriptors for user's intention and items. What important about these two sources is that they can be found with visualizations. People would describe the charts and their insights to a reader. So, they can be extracted without asking the user to explicitly provide them. Yet, to investigate how effectively they encode information, each information source (tags, titles, Q&As) is characterized using information-theoretic measures, such as entropy, conditional entropy and mutual information, as suggested in Chi et al. [CM08, LTBH15]. Subsequently, using each of these sources separately, we build models for user and item profiles to recommend personalized visualizations applying a content based recommender. However, our visual recommender includes both collaborative and content-based recommender. For the sake of completeness, we finally used those models for our hybrid recommender to define personalized visualizations (see Figure 6.1). This allows us to obtain insights and draw general conclusions about the drawbacks and benefits for each source as input for the visual recommender systems. This data was obtained with the crowd-sourced study presented in Chapter 4.

In a nutshell this Chapter makes the following contributions:

- It introduces a framework to assess the encoding power of different textual information sources in describing user preferences and visualizations.
- The framework is used in a thorough analysis of different kinds of user-provided input characterizing data models for user and visualizations.
- We derive insights on how their nature impacts the generation of personalized visualization recommendations.

Note, the following sections are based on the publication P8 [MVT17].

## 6.1 Related Work

One of the key concerns in personalizing recommendations is building personalized profiles of individual users and candidate items. These profiles constitute models of:

1. individual user characteristics describing what the user needs and prefers—user model
2. item characteristics describing what the items represent, their content respectively—item model.

Yet content-based recommender systems try to define personalized recommendations by matching up the attributes of the user model with the attributes of the item model. However the following questions arise:

1. which source of information is most effective at encoding user preferences and item characteristics?
2. which source of information yields the more accurate recommendations?
3. how to acquire this information from the user?

In traditional content based recommendation approaches, systems collect the user preferences by explicitly asking the users to share their interest and needs, typically in form of tags. Although partially successful, these approaches often suffer on the missing motivation of the user for annotations [FSCG07, LZMT15]. However, recent studies on this topic show that user's motivation to annotate resources increases if this provides a navigational aid to the resources [SKK10]. Ricci et al [RN07], for instance, present a recommender system to help the user with searching for travel products. To define recommendations that are closer to user's needs, the system asks the user to provide critiques in form of textual feedback when one feature of the recommended product is not satisfactory or very important. The authors prove the effectiveness of their system with an empirical study. This also applies for visualizations. When the user annotates visualizations, she provides her insights and her interpretation on the data being visualized. Hence, the annotations serve as analysis finding records and personal reminder for later data discovering and analysis tasks [EB12].

The process of annotating can be considered as an encoding process where the annotations encode the information (facts, features etc.) about the items [SKK10]. However, it depends on the encoding quality of the used annotation type (tags, titles, Q&As) how good a recommender performs. Chi et al. [CM08] use information-theoretic measures (entropy, conditional entropy, mutual information) to evaluate the encoding power of tags collected from the social tagging site *del.icio.us*. Using these measures, Chi et al. quantify the diversity in tags and documents and

the amount of shared information between them. The obtained results finally provide insight into how effective the tags are at encoding documents. Strohmaier et al. [SKK10] use conditional entropy and orphan ratio for measuring and detecting the tacit nature of tagging motivation by analyzing the tag sets produced by 8 different tagging systems regarding to their encoding and descriptive power. The results of their study show that (i) tagging motivation of individuals varies within and across tagging systems and (ii) user’s motivation for tagging has an influence on produced tags and folksonomies. Yi-ling Lin et al. [LTBH15] analyze the tags they collected for images in two different tagging conditions (with and without description) on perspectives such as diversity, specificity, quality, similarity and descriptiveness. The analysis mainly covers common text-quality metrics such as number of unique- and common words of each content and so.

Regarding on the research question “which source of information defines the more accurate recommendations” the most notable research is provided by Bellogín [BCC10] et al., who try to identify the sources of information (ratings, tags, social contacts, etc.) most valuable for a recommender in a social music service. To do so, Bellogín et al. evaluate a number of content based, collaborative filtering and social recommenders on heterogeneous datasets obtained from Last.fm with well-known metrics precision, recall and ranking based matrix. Next, they compare the characteristics of the generated recommendations using non-performance metrics such as coverage, overlap, diversity and novelty between different set of recommendations.

Our work extends the research on evaluating the power of annotations (tags, ratings etc.) at encoding documents, music tracks and users’ interests in these resources into encoding visualizations and users’ visualization preferences. Similar to the relevant works, we suggest for this purpose information-theoretic measures entropy, conditional entropy and mutual information as these measures have been proposed in many fields to assess the diversity of textual content [Yao03, JM04, MC08]. Furthermore, we address the question of how valuable different source of information are for recommending visualizations by applying a content based filtering approach as this recommendation approach builds on the content features and a hybrid approach which includes both content-based and collaborative filtering approaches.

## 6.2 Experiment Setup

The goal of the study is to investigate the characteristics of tags, titles and Q&As and their impact on recommending personalized visualizations. To collect these different kinds of annotations we designed a crowd-sourced study where we asked the user to annotate and rate the visualizations according to the different data sources. The details about how we collected the annotations (tags, titles, Q&As) can be taken from Chapter 4. In Section 4.2.2, we have already investigated the characteristics of the collected ratings and their impact on the recommendation quality— on our

CF-RS respectively (see also [MVTS15a]). Thus, here we put focus on tags, titles and Q&As. To that end, we proceed with the experiment as follows:

- First, we analyze how good these three types of input data models encode both user and visualizations (see Section 6.3). The observations from this part of the experiment shall reveal important facts about why some of the inputs are better than others. Based on those observations, we build a list of candidate inputs for each data set (i.e., which descriptors accurately describe the user and which ones the visualizations). Those are in the end the assumptions that we want to confirm using the offline study.
- Next, we execute our content-based and hybrid recommender on candidate input data models to see if their encoding power can be confirmed. The results of this study are presented in Section 6.4.

## 6.3 Encoding power of User-provided Input

In this study we aim to explore the characteristics of different user-provided input (annotations) in terms of encoding the users and the visualizations. Information-theoretic measures are used to characterize the tags, titles and Q&As.

### 6.3.1 Methodology

For the analysis of tags, titles and Q&As we employ information-theoretic measures: entropy, conditional entropy and mutual information. Using information-theoretic measures, we are able to:

- quantify the diversity in annotations (terms in further text), their encoding power respectively,
- the amount of shared information between terms describing the users and the items (visualizations).

With this information, we expect to answer why one input might be more suitable for recommending visualizations than the others.

In information theory, entropy measures the amount of uncertainty in a single random variable [Gra90]. Given a random variable ( $X$ ), which consist of occurrences  $\{x_1 \cdots x_N\}$ , each of which occurs with the probability  $p(x)$ , the entropy  $H(X)$  is defined as:

$$H(X) = - \sum_{x \in X} p(x) \log(p(x))$$



**Table 6.1:** Basic statistical properties of the datasets collected via Amazon Mechanical Turk. Column "User/Vis" shows the average number of the user assigned to a visualization; "Vis/User" is the average number of visualizations assigned to a user. Note that the values in parenthesis indicate number of unique terms.

Dataset	#Vis	#User	Users/Vis	Vis/User	#Tags	#Titles	#Q&As
Movies	55	36	10	15.27	2731 (292)	2217 (295)	2638 (822)
EU	30	19	10	15.79	1403 (166)	1394 (234)	1354 (514)
Books	7	15	10	4.67	349 (87)	270 (92)	395 (188)

Conditional entropy [Gra90], on the other side, measures the uncertainty in a random variable given the value of another random variable. Given two discrete random variables  $X = \{x_1 \cdots x_N\}$  and  $Y = \{y_1 \cdots y_N\}$  so that the event  $(x, y)$  occurs with the joint probability  $p(x, y)$ , the joint entropy is defined as:

$$H(Y, X) = - \sum_{\{y,x\} \in \{Y,X\}} p(y,x) \log(p(y,x))$$

Using this value, conditional entropy is defined as  $H(Y|X)$  [CM08]:

$$H(Y|X) = H(Y, X) - H(X)$$

Concretely, conditional entropy quantifies the amount of information needed to describe the variable  $X$  (e.g., the user or the visualization) when the value of the variable  $Y$  (e.g., tags, titles, Q&As) is known. If  $H(Y|X)$  is minimized, each tag (or title, Q&A) uniquely refers to an individual user (or visualization) [CM08]. In contrast, when  $H(Y|X)$  is maximized, each tag (or title, Q&A) is as likely as all others.

Finally, mutual information [Gra90] is a measure of independence between two random variables. In other words, it quantifies the amount of data (information) shared (mutual) between variables. Given two discrete random variables  $X = \{x_1 \cdots x_N\}$  and  $Y = \{y_1 \cdots y_N\}$  so that the event  $(x, y)$  occurs with the joint probability  $p(x, y)$ , the mutual information  $I(X; Y)$  is defined as:

$$I(X; Y) = H(Y) - H(Y|X)$$

High mutual information indicates a large dependency between two variables. In contrast, if the mutual information is minimized the variables are independent.

### 6.3.2 Results

Table 6.1 summarizes basic statistics for tags, titles and Q&As and shows the distribution of the entire terms<sup>1</sup> over the user and the visualizations. As introduced

<sup>1</sup>A term is considered here as a single word e.g., in a tag input data model, a term corresponds to a single tag.

**Table 6.2:** Example distributions of top-5 terms for the EU dataset. Note, the terms are stemmed using Porter stemmer [KG13].

<b>Tags</b>	<b>Count</b>	
	# of visualizations	# of user
constitut	15	13
educ	16	11
monarch	15	16
republ	15	15
popul	27	18
<b>Titles</b>		
constit	15	8
countri	22	5
european	14	8
popul	21	8
valu	17	8
<b>Q&amp;As</b>		
inform	14	10
larg	16	9
onlin	13	9
republ	13	10
type	13	10

earlier, each of the visualizations in a particular dataset was individually evaluated regarding to tags, titles and Q&As, i.e., 55 visualizations in Movies dataset, 30 and 7 in EU and Books respectively (cf. the second column in Table 6.1). For this configuration, the user involved in the study have provided overall 4483 tags, 3881 titles and 4387 Q&As (2% yes/no Q&As). The average word length (char) was 5.2 for tags and questions, 5.3 for titles and 4.7 for answers. An excerpt of the most popular terms for EU is shown in Table 6.2. Some important differences between collected data could already be identified when considering this distribution in conjunction with unique terms (Note that the number on unique terms are enclosed with parenthesis, see Table 6.1). According to descriptive data from the table, 10.69% of the tags, 13.31% of titles and 31.16% of Q&As were unique, i.e., not globally repeated. The fact that a question typically associates with only one specific visualization may explain this phenomenon. Taking this cue, we can assume that the varied number on different type of terms directly affects the recommendation quality. In brief, the more unique terms are applied to a visualization the easier it should become to discriminate this visualization in the finding process from others. Subsequently, the more individual terms a user provides, the higher the ability should be to accurately direct this user to the preferred visualizations [SKK10]. However, the more accurate way to measure how good a term is in discriminating a resource from others is measuring the value of the information it provides about a resource and about

**Table 6.3:** Information-theoretic measures for tags, titles and questions used for the user profiles. Note that the measures have been calculated among all three datasets.

Datasets	Term	User Model		
		Entropy	Conditional Entropy	Mutual Inf.
Movies, EU, Books	Tags	5.9376	3.0381	2.8995
	Titles	6.1421	2.9815	3.1606
	Questions	6.8898	3.1436	3.7462

the user. For this purposes we investigate in the following, first, the power of users' terms at encoding users' visual preferences and, next, at encoding the content of visualizations.

### 6.3.2.1 Power of user-provided input at encoding users

To investigate the quality of extracted terms at encoding users' visual preferences we applied the information-theoretic measures among all three datasets. Considering all three datasets in our analysis helps us to achieve more objective results, compared to analyzing each dataset individually.

In this experiment,  $X$  is the users and  $Y$  is either the tags, the titles or the Q&As. The analysis intends to determine which of  $H(Tags)$ ,  $H(Titles)$ ,  $H(Q&As)$  indicates more diversity, which of  $H(User|Tags)$ ,  $H(User|Titles)$ ,  $H(User|Q&As)$  has more power in describing the users and which of  $I(User; Tags)$ ,  $I(User; Titles)$ ,  $I(User; Q&As)$  has higher value and can specify the users better. Table 6.3 summarizes the results of this study. Note, to follow a common design principle of interactive (question-answering) systems, we suggest to split the Q&As input so that questions are used for the user- and answers for the item model.

When considering the results in Table 6.3, at the first look we can observe that the entropy ( $H(Questions)$ ) is higher than ( $H(Titles)$ ) and ( $H(Tags)$ ). This suggests, the users provided more diverse and specific questions than titles and tags. Given this fact, we hypothesize that questions have a strong encoding power. Yet, entropy measures the amount of uncertainty. Conditional entropy, however, quantifies the amount of uncertainty in a random variable (i.e., the user) given the value of another random variable (i.e., tags, titles or questions).

We therefore consider next the entropy of the users conditional on tags (or titles, questions), i.e.,  $H(User|Tags)$ ,  $H(User|Titles)$  and  $H(User|Questions)$  (see Table 6.3 second column). Looking at the results,  $H(User|Questions) > H(User|Tags) > H(User|Titles)$ . What that means is, that tags and titles have a strong power in describing the user than the questions.

Yet, conditional entropy is a relative measure and tells little about the independence between the tags (or titles, questions) and the user [CM08]. The independence, however, matters in recommender systems when it comes to defin-

**Table 6.4:** Information-theoretic measures for tags, titles and answers used for item profiles. Note that the measures have been calculated among all three datasets.

Datasets	Term	Item Model		
		Entropy	Conditional Entropy	Mutual Inf.
Movies, EU, Books	Tags	5.9376	4.1429	1.7947
	Titles	6.1421	4.1384	2.0037
	Answers	6.6371	2.7405	3.8966

ing a link between the user and the resources. Thus, to complete the analysis on tags (or titles, questions) in specifying the user, we next, analyze the amount of information shared between tags (or titles, questions) and a user, mutual information ( $I(User; Tags)$ ,  $I(User; Titles)$ ,  $I(User; Questions)$ ) respectively. The results show that  $I(User; Questions)$  is the highest compared to  $I(User; Tags)$  and  $I(User; Titles)$  (see Table 6.3 last column). Yet, these results finally suggest, questions are more effective in specifying the user than the tags and the titles.

### 6.3.2.2 Power of user-provided input at encoding visualizations

Similar to our previous study, to investigate the general quality of the user-provided input at encoding visualizations we applied the information-theoretic measures among all three datasets. In this case,  $X$  are visualizations and  $Y$  are either tags, titles or answers. The analysis intends to determine which of  $H(Tags)$ ,  $H(Titles)$ ,  $H(Answers)$  indicates more diversity across visualizations, which of  $H(Vis|Tags)$ ,  $H(Vis|Titles)$ ,  $H(Vis|Answers)$  has more power in describing visualizations and which of  $I(Vis; Tags)$ ,  $I(Vis; Titles)$ ,  $I(Vis; Answers)$  has higher value and can specify visualizations better. Table 6.4 summarizes the results of this study.

The Table 6.4 reveals that the entropy of answers is higher than of tags and titles. At a first glance, this indicates, the visualizations have been annotated with more specific and unique answers than tags and titles. However, as we noted in the previous study, entropy just measures the amount of uncertainty in a random variable (i.e., visualization) given the value of another random variable (i.e., tags, titles or answers). When considering  $H(Vis|Tags)$ ,  $H(Vis|Titles)$ ,  $H(Vis|Answers)$ , we observe that answers are more unique and special than tags and titles ( $H(Vis|Answers) < H(Vis|Tags)$ ,  $H(Vis|Answers) < H(Vis|Titles)$ ) (see Table 6.4 second column). Thus, it might be more difficult for the system to retrieve a visualization that has been annotated with a certain tag or title than with a certain answer. To validate this we finally measure the degree of independence between tags (or titles, answers) and a visualization—the amount of information shared (mutual)  $I(Vis; Tags)$ ,  $I(Vis; Titles)$ ,  $I(Vis; Answers)$ . Remember, full independence is reached when e.g.,  $I(Vis; Tags)$  is zero.

Table 6.4 (last column) shows the mutual information  $I(Vis; Answers)$  is higher

than of  $I(Vis; Tags)$  and  $I(Vis; Titles)$ . These results, finally, suggest a high quality of answers at encoding visualizations. Taking this cue, we can assume the answers are powerful to direct the user to the corresponding visualizations than tags and titles.

### 6.3.2.3 Summary

Using information-theoretic measures we aimed to characterize tags, titles and Q&As in describing the user and the items (visualizations). To that end we performed two studies where we analyzed the power of (i) tags, titles and questions at encoding the user and (ii) the tags, the titles and the answers at encoding visualizations. The findings of the studies should help in predicting performance of the potential candidates for the user- and item models being used for our visual recommender.

Results suggest a strong link (dependency) between the user and her questions and the items and their (assigned) answers. This assumption is made regarding to the shared information between (i) the user & questions and (ii) the item & answers,  $I(User; Questions)$ ,  $I(Vis; Answers)$  respectively. Namely, the results of  $I(User; Questions)$ ,  $I(Vis; Answers)$  show that a set of specific terms from questions refers to an individual user and each answer to a specific item. Yet, this is an essential finding for designer of content-based recommender systems. It suggests using questions for the user modeling and answers for the item modeling.

To verify this assumption, we build, in the following, the user and the item models using user's questions and answers and explore the quality of the generated recommendations in an offline study employing our CB based recommender system. We applied this recommender technique since it is traditionally used for the user-provided input, such as tags, comments, etc. We measured the quality of the recommendations by their closeness to what the user prefers and needs.

Note, for the sake of completeness, we also included additional setting where tags are taken for the user- and the item models. Considering the results in Section 6.3.2.1, the quality of the generated recommendations should be lower when using this combinations, since tags have a lower mutual information than Q&As. Moreover, to verify the low performance of titles, settings with titles are reported too.

In the following we describe the method and metrics used to validate our approach in detail and present the results of the offline study.

## 6.4 Recommendation Quality

In this section, we summarize the results of the offline evaluation. As defined in Section 6.2, first, we show how VizRec's content-based recommender performs when exploiting alternative input sources (titles, Q&As) collected in Amazon Turk experiment. We start with CB as this recommendation approach builds on the content features. However, VizRec is a hybrid recommender. For the sake of completeness,

we next analyze the performance of our hybrid recommender when exploiting user’s titles and Q&As with their ratings.

### 6.4.1 Content Based Recommender

This section details the offline evaluation we performed to measure the recommendation quality of our content-based (CB) recommender when exploiting alternative information sources (e.g., titles, Q&As), beside commonly used tags.

#### 6.4.1.1 Methodology

Following the method described in [MVT16], we split the preference model including either users’ tags, titles or questions into the two distinct sets, one for training the recommender (training-set) and another one for testing (test-set). The test-set acts here as a reference value that, in an ideal case, has to be fully predicted for the given training-set. From each of the datasets in the preference model, we randomly select 80% of user’s data and enter them into the training-set performing five-fold cross validation. The recommendations produced out of the training-set are further used to evaluate the performance of our recommender. The performance of the recommender depends generally on how good it predicts the test-set. We compared the generated recommendations (prediction-set) and the test-set by applying a variety of well-known evaluation metrics in information retrieval [HKTR04]: Recall (R), Precision (P), F-Measure (F), Mean Average Precision (MAP) and the Normalized Discounted Cumulative Gain (nDCG). The first three metrics basically express the quantity of relevant recommended results, whereas MAP and nDCG quantify the concrete ordering of the results (i.e., penalizing results which are not on the top but are relevant for the user). We refer to the research papers [Rij74, RKAES13, PS13] and the Chapter 5 for more detailed definitions of the evaluation metrics. Note, the tests are performed for each user- and item model combination independently.

#### 6.4.1.2 Results

To measure the improvements in terms of recommender quality (=accuracy, relevance), we compared the individual CBs ( $CB_{Tags,Tags}$ ,  $CB_{Titles,Titles}$ ,  $CB_{Q,A}$ ) with the baseline filtering algorithm Random (RD). The RD method simulates the recommender behavior providing a random rating for each visualization. Note, for the Q&As based CB approach ( $CB_{Q,A}$ ) we used user’s questions in user- and user’s answers in item model.

For the comparison, we analyzed the top 3 recommendations ( $k=3$ ), since our datasets are relatively smaller than some commonly used datasets, such as CiteULike and BibSonomy. Table 6.5 shows the quality metrics values  $F@3$ ,  $MAP@3$ ,  $nDCG@3$  estimated for the three datasets.

Yet, when considering the recommendation accuracy ( $F@3$ ), at a first glance, we can observe that tags based CB ( $CB_{Tags,Tags}$ ) outperforms for all three datasets

**Table 6.5:** The performance of our individual content based filtering approaches (CB), compared with baseline algorithm RD: quality metric values considering the first three recommendations in the list (k=3). \*\*\*Significant at  $p < 0.001$ .

Dataset	Algorithms	Metric		
		F@3	MAP@3	nDCG@3
Movies	RD	0.0055	0.0020	0.0048
	CB <sub>Tags,Tags</sub>	0.0740***	0.0545***	0.0830***
	CB <sub>Titles,Titles</sub>	0.0650***	0.0500***	0.0743***
	CB <sub>Q,A</sub>	0.0547***	0.0450***	0.0643***
EU	RD	0.0150	0.0044	0.0103
	CB <sub>Tags,Tags</sub>	0.1862***	0.1120***	0.1801***
	CB <sub>Titles,Titles</sub>	0.1726***	0.1030***	0.1663***
	CB <sub>Q,A</sub>	0.1505***	0.1014***	0.1642***
Books	RD	0.0333	0.0333	0.0420
	CB <sub>Tags,Tags</sub>	0.2360***	0.2077***	0.2700***
	CB <sub>Titles,Titles</sub>	0.2310***	0.2133***	0.2720***
	CB <sub>Q,A</sub>	0.2267***	0.2233***	0.2720***

the baseline algorithms RD (cf.  $F@3(CB_{Tags,Tags}) = 0.0740$ ,  $F@3(RD) = 0.0055$  for Movies). So, we hypothesize that the experimentation with individual user- and item models has had some effect among all three datasets. To discover what the effect was and how significant it is, we performed statistical tests which we report in the following.

The results for F@3, MAP@3, nDCG@3 have been analyzed independently for each dataset applying Friedman’s ANOVA. Note, we used this test since our data were not normally distributed and (per dataset) the same participants have been used for each individual CB approach. The results for all three datasets show a significant effect of the used type of the item- and the user models on the recommendation accuracy (F@3), with  $\chi^2(4) = 25.10$  for Movies,  $\chi^2(4) = 19.80$  for EU and  $\chi^2(4) = 20.14$  for Books,  $p < 0.001$ . To explore where the differences lie we applied *Post hoc* tests (multiple comparisons between groups) with Bonferroni correction. The results for all three datasets reveal a significant difference between the values of the individual CB approaches (CB<sub>Tags,Tags</sub>, CB<sub>Titles,Titles</sub>, CB<sub>Q,A</sub>) and baseline algorithm. Note, the critical difference ( $\alpha = 0.05$  corrected for the number of tests) was 28.10 for Movies, 20.10 for EU and 18.65 for Books. However, there were no significant differences between the values of individual recommenders (CB<sub>Tags,Tags</sub>, CB<sub>Titles,Titles</sub>, CB<sub>Q,A</sub>),  $p > 0.05$ .

Looking at the results for MAP@3 and nDCG@3 measures which examine the ranking of the recommended visualizations we observe similar results. Concretely, the results show a significant effect of the used type of the item- and the user models on the ranking of the recommendations, with  $\chi^2(4) = 23.56$  for Movies,  $\chi^2(4) = 20.10$  for EU and  $\chi^2(4) = 18.65$  for Books,  $p < 0.001$ . Similar to the previous analysis, to explore where the differences lie we applied *Post hoc* (multiple

comparisons between groups) tests with Bonferroni correction. The results of *Post hoc* tests shown for all three datasets, when tags, title and Q&As based models are used, the visual recommender can sort the recommendations according to their relevance better than baseline algorithm. The critical difference ( $\alpha = 0.05$  corrected for the number of tests) was 28.10 for Movies, 20.10 for EU and 18.65 for Books. The results for nDCG confirmed the results we obtained for MAP@3 measures showing a significant improvement by ranking of recommendations when using either (i) tags, (ii) titles or (ii) Q&As based models compared to the random baseline algorithm.

## 6.4.2 Hybrid Recommender

This section details the offline evaluation we performed to measure the recommendation quality of our hybrid (HB) recommender when using the user ratings with alternative information sources (e.g., titles, Q&As), beside commonly used tags.

### 6.4.2.1 Methodology

To evaluate the performance of our hybrid recommender, we used the same quality metrics as introduced in Section 6.4.1. We run the hybrid recommender for each user- and item model (tags, titles, Q&As) separately using a 5-fold cross validation and compared its performance with the baseline algorithms MostPopular (MP). We chose MP as the baseline algorithm since it generates the recommendations sorted according to global ratings, in our case accumulated from ratings of individual users. Given that, we use the user ratings in our hybrid recommender (for the CF part respectively), it is more accurate to apply MP instead of RD. The results of this experiment are summarized in Table 6.6.

### 6.4.2.2 Results

When considering the recommendation accuracy (F@3), at a first glance, we can observe that tag based hybrid ( $HB_{Tags}$ ) outperforms for all three datasets the baseline algorithms MP. Moreover, for the datasets Movies and EU the tag based hybrid outperforms title ( $HB_{Title}$ ) and Q&As ( $HB_{Q,A}$ ) based hybrids. Interestingly, for the dataset Books the amount of the accuracy for  $HB_{Tags}$  and  $HB_{Title}$  are the same and higher than for  $HB_{Q,A}$ . So we again hypothesize that the experimentation with different datasets and data models has had some effect. To discover what the effect was and how significant it is, we performed statistical tests which we report in the following.

Again, the results for F@3, MAP@3, nDCG@3 have been analyzed independently for each dataset applying Friedman’s ANOVA. The results for all three datasets show a significant effect of the used type of item- and the user models on the recommendation accuracy (F@3), with  $\chi^2(4) = 41.51$  for Movies,  $\chi^2(4) = 30.63$  for EU and  $\chi^2(4) = 27.12$  for Books,  $p < 0.001$ . To explore where the differences lie, we applied *Post hoc* (multiple comparisons between groups) tests with Bonferroni correction.



**Table 6.6:** The performance of our individual hybrid filtering approaches (HB), compared with baseline algorithm MP: quality metric values considering the first three recommendations in the list (k=3). \*\*\*Significant at  $p < 0.001$ .

Dataset	Algorithms	Metric		
		F@3	MAP@3	nDCG@3
Movies	MP	0.0591	0.0163	0.419
	HB <sub>Tags,Tags</sub>	0.3246***	0.3988***	0.4640***
	HB <sub>Titles,Titles</sub>	0.2615***	0.2796***	0.3341***
	HB <sub>Q,A</sub>	0.2725***	0.2935***	0.3581***
EU	MP	0.0211	0.0088	0.0161
	HB <sub>Tags,Tags</sub>	0.5444***	0.5303***	0.6073***
	HB <sub>Titles,Titles</sub>	0.4406***	0.4191***	0.4997***
	HB <sub>Q,A</sub>	0.3623***	0.3057***	0.3876***
Books	MP	0.0667	0.0444	0.0667
	HB <sub>Tags,Tags</sub>	0.5680***	0.8600***	0.8615***
	HB <sub>Titles,Titles</sub>	0.5680***	0.8600***	0.8615***
	HB <sub>Q,A</sub>	0.5120***	0.7566***	0.7600***

The results for all three datasets reveal a significant difference between the values of the individual HB approaches (HB<sub>Tags</sub>, HB<sub>Titles</sub>, HB<sub>Q,A</sub>) and baseline algorithm. Note, the critical difference ( $\alpha = 0.05$  corrected for the number of tests) was 28.10 for Movies, 20.10 for EU and 18.65 for Books. However, there were no significant differences between the values of individual recommenders (HB<sub>Tags</sub>, HB<sub>Titles</sub>, HB<sub>Q,A</sub>),  $p > 0.05$ .

Yet, for MAP@3 and nDCG@3 measures we observe similar results. Concretely, the results show a significant effect of the used type of the item- and the user models on the ranking of the recommendations, with  $\chi^2(4) = 50.60$  for Movies,  $\chi^2(4) = 39.92$  for EU and  $\chi^2(4) = 32.10$  for Books,  $p < 0.001$ . Similar to the previous analysis, to explore where the differences lie we applied *Post hoc* (multiple comparisons between groups) tests with Bonferroni correction. The results of *Post hoc* tests shown for all three datasets, when tags, title and Q&As based models are used, the visual recommender can sort the recommendations according to their relevance better than the baseline algorithm. The critical difference ( $\alpha = 0.05$  corrected for the number of tests) was 28.10 for Movies, 20.10 for EU and 18.65 for Books. The results for nDCG confirmed the results we obtained for MAP@3 measures showing a significant improvement by ranking of recommendations when using either (i) tags, (ii) titles or (ii) Q&As based models compared to the MostPopularr baseline algorithm.

### 6.4.3 Discussion

The main outcome of our study is that all three inputs (tags, title and Q&As) show a comparable quality in recommending visualizations. This result is important because it gives the designer freedom in choosing the method for preference elicitation.

Besides, it makes the suggested approach applicable in domains in which only particular types of inputs can be supported (e.g., question-answering systems).

We could confirm this result for all datasets, as illustrated in Table 6.5. Moreover, all three inputs are, as expected, significantly better at encoding visualizations than the baseline algorithms RD and MP. Also, when considering the results in more detail, i.e., the quality  $F@3$  and the sorting accuracy ( $MAP@3$  and  $nDCG@3$ ), it does not matter which of the inputs to use. (Note that there are negligible differences in means, which are statistically not significant). This would, in the end, mean that characteristics of the individual inputs are very close to each other. In fact, providing a title would be nothing else but providing a set of tags (in terms of how many and which words have been provided). We analyzed these characteristics in the first part of our study.

Using information-theoretic measures we found that some inputs better encode the user/visualizations than the other. In particular, questions and answers have been identified to show distinctive characteristics compared to tags and titles. It turned out that they more precisely address a particular user/visualization, since, as results reveal, they have terms which are less common (shared) than in the case of tags and titles. This, in fact, comes from the nature on how questions/answers are built. For instance, it is more likely that similar or same words are provided when describing visualizations via tags rather than using complex sentences. Generally, the users are familiar when describing resources in form of tags, as tagging approach is quite intuitive and straightforward. Using question/answers, instead, is more subjective. One aspect here is building a sequence of words (a sentence) and another is using proper adjectives in that sequence. These terms also contribute to the user/item model. Nevertheless, as shown later in the offline experiments, these differences were not significant enough to be manifested by the content-based recommender (at least with the cosine similarity metric we chose) and in turn in hybrid recommender.

## 6.5 Summary

In this Chapter we investigated the power of different kinds of user-provided input to effectively encode user’s visual preferences and the content of visualizations. To do so we employed information-theoretic measures including entropy, conditional entropy and mutual information. Using these measures, we were able to quantify the diversity in individual inputs, their encoding power respectively and also the amount of shared information between them and the users/visualizations. The outcome of the study should suggest a list of potential candidates to build the user models defining users’ interest/needs and item models describing the content of the visualization— both crucial for content-based recommender systems. Finally, we executed our content-based and hybrid recommender on candidate models to see if their encoding power could be confirmed. In other words, we performed an offline

study to assess the practicability of the individual models in recommending personalized visualizations. The data we used for the above studies was collected in the scope of the empirical study, where we involved 47 participants to annotate different types of visualizations using tags, titles, questions and answers. More details about the study are given in Chapter 4 and Section 8.2.

Regarding to our first study, we found that the best user- item model combination is guaranteed when using questions for the user- and answers for the item models (considering their mutual information values). The offline study has confirmed the good quality of this combination as it produced better recommendations than the baseline algorithm. However, the recommendation results for this combination was not significantly better or different than that of the tags and titles. Although differences at encoding power between the individual inputs could be manifested, those differences were negligible and not crucial for the content-based and hybrid recommender system we employed. Nevertheless, the fact that the recommendation quality and accuracy were still high using the alternative inputs, titles and Q&As respectively, demonstrated the capability of these inputs being used for visual recommender systems.

In summary, this Chapter shows the good quality of alternative input types (titles, Q&As) to derive high quality visualization recommendations. It further emphasizes the relevance of annotations for the users as they directly link them to the items which might be closer to what they need and prefer.

# Chapter 7

## Conclusion

This thesis is motivated by the question, how to provide meaningful and sophisticated methodologies to automate and personalize the visualization process to better reflect user’s preferences, needs and interest. To address this question, the thesis proposes methods that comprise different research areas. Information visualization focuses on representing data in a way that makes intelligent use of human visual abilities. Recommender systems, however, support the user in decision-making process and provide the user with items that are closer to what she needs and prefers. Together, these two research areas define the building blocks of our visual recommender which is able to, first, automatically define a list of appropriate visualizations considering visual encoding rules and perceptual guidelines, second, filters a subset considering user’s visual preferences and topic of interest. Our approach to generating and recommending personalized visualizations, the process of elicitation of users’ preferences and the insights described in this thesis are to the best of our knowledge, novel. The developed methods followed an extensive analysis of the existing research, one lab study, one crowd sourced study and six offline studies.

Finally, the thesis resulted in eight peer-reviewed scientific research publications. They consist of two journal papers, four conference papers, one demo and one poster.

### 7.1 Research Question

To achieve the goal of the thesis, we defined three research questions. The answers to these questions define the main contributions of this thesis. The questions and the corresponding answers are provided in the following:

**RQ1: Which aspects of the visualizations have to be considered to automatically generate and recommend visualizations for non expert user across various fields/domains?**

Research Question 1 is answered in Chapter 3 and publications P1 [MHS<sup>+</sup>13], P2 [MHT<sup>+</sup>14] and P3 [STV<sup>+</sup>14]. In Chapter 3 and our scientific publications, we

propose a vocabulary (Visual Analytics Vocabulary) which explicitly conceptualizes visualizations using pragmatic and simple facts. This includes formulating visual properties of the visualizations, their concept, characteristics and functions given to them. The theoretical underpinning for our vocabulary is presented in Chapter 2. The concepts described in Chapter 2 had a vast impact on the vocabulary as it provides expert knowledge about general rules and principles in visual encoding and serves as a guideline for modeling visualizations. The literature review presented in Chapter 2, however, defined the frame for visualization modeling and also showed what is missing in this field of research. Considering the main limitations in existing research, our proposed visualization vocabulary makes the following contributions:

- It describes the relation between the visualizations and visual properties, provides semantics to the visualizations and their components
- It comprises multidimensional visualizations
- It is extensible with new visualizations
- It is domain- and technology independent
- It is machine- and human readable

As a building block, the VA Vocabulary is applied to define an OWL ontology of visualizations. Combined with a specialized algorithm (mapping algorithm) the ontology is used in a rule-based recommender system to automatically define the list of appropriate visualizations for user's data. In summary, this thesis has shown that automated visualization process for non expert user across various domains can be realized, when (i) using a unique, domain-and technology independent description of visualizations (including multidimensional visualizations) which can be extended with new (e.g., domain specific) visualizations and (ii) applying a mapping algorithm which is able to determine a valid path between data and visualizations by verifying the datatype and structural compatibility between them.

Together with the mapping algorithm, the Visual Analytics Vocabulary, has been successfully applied in the following platforms: (1) CODE-Platform to automatically visualize research data embedded in scientific publication, and (2) EEXCESS-Platform to automatically define appropriate visualizations for the recommended items from scientific and educational content (i.e., papers, images, videos).

**RQ2: Can standard recommender system strategies be exploited to personalize the visualization recommendation in order to better address user preferences/needs?**

Research Question 2 is answered in Chapter 5 and publications P4 [VMdS<sup>+</sup>15], P5 [MVTS15b], P6 [MVTS15a] and P7 [MVT16].

It requires specific expertise in information visualizations and underlying data to determine the right visual representation for the underlying data. Although there are rules (applied in our Visual Analytics Vocabulary) that help generate them, the results are too broad to account for varying user preferences. To tackle this issue, we proposed a novel recommender system that suggests visualizations based on a set of visual cognition rules and personalize them considering user's preferences. Our goal was to recommend those visualizations that are closer to what the user needs and prefers.

We investigated different strategies to recommend personalized visualizations considering different aspects of user preferences. First, we investigated a Collaborative filtering (CF) approach which is build on the user ratings. Content-based (CB) filtering is analyzed as a means to suggest visualizations according to the information needs of the user by collecting vectors of tags describing what aspects of the data the user is interested in. To benefit of the strength of CF and CB, we build a hybrid recommender which is a combination of both aspects, what a visualization is about (tags), and how good it is (ratings). Furthermore, an extensive evaluation of visualization types in the context of three data repositories was conducted in Amazon Mechanical Turk. The evaluation served multiple purposes. The first goal was to elicit user preferences (ratings, tags etc.) for visualizations. The second goal was to assess and understand quality and variability of visualizations whereas the last goal was to collect user data to generate personalized recommendations. Yet, the collected user preferences not only provided the evidence that the user preferences vary widely for visual representations generated automatically and only a personalized approach would accommodate user's visual preferences and needs better. They also have proven that popular recommendation strategies (CF, CB and Hybrid) can be successfully applied in visual recommender systems to personalize visualizations.

The personalized visual recommender is currently used in the AFEL-platform as a personalized learning dashboard for monitoring and assessing multi-dimensional statistical information collected during learning activity (e.g. learning progress, view frequencies etc.).

**RQ3: Which methods for preference elicitation to use to define personalized visualizations that are closer to user's needs and interests?**

Research Question 3 is answered in Chapter 6 and publication P8 [MVT17]. This research question concentrated on characterizing the user preferences, in particular the sources of information used to describe user's visual preferences and interest and the content of the visualizations. We considered three sources corresponding to different aspects of interest: a title that describes the chart, a question that can be answered with the chart (and the answer), and a collection of tags describing features of the chart. We investigated the user-provided input based on these sources collected with a crowd-sourced study. Firstly, information-theoretic measures are applied to each source to determine the efficiency of the input in describing user

preferences and visualization contents (user and item models). Secondly, the practicability of each input is evaluated with content-based recommender system and as well as with the hybrid recommender. The findings of this investigations highlighted that all three inputs (tags, title and Q&As) show a comparable quality in recommending visualizations. What this means is, the designer of visual recommender systems can choose between three methods (tagging, titling, question-answering) as all of them produced recommendations equal in quality. Yet, the overall methodology and results contribute methods for design and analysis of visual recommender systems.

Considering the insight we gained during addressing Research Question 3, the current version of our personalized visual recommender is build upon user's tags and ratings.

## 7.2 Self- Reflection about this Thesis: Lesson Learned

This section provides a self-reflection about the development of the thesis. It serves as a summary of experiences collected during designing, developing and evaluating the personalized visual recommender and summarizes the most important insights gained.

The experiences we made and the insights we gained with regard to the Visual Analytics Vocabulary, one of the building blocks of the rule-based recommender, can be summarized as follows:

- The development of a visualization vocabulary requires background knowledge about visual encoding rules and perceptual guidelines. To achieve this, we studied the existing literature about the principles of information visualization. Concretely, we put focus on Bertin's work on the theory of information design and Mackinlay's formal language for generating graphical representations. Given that the whole knowledge about the visualization domain is too extensive to be completely covered in one vocabulary, we decided to cover as many related concepts as possible in our vocabulary in order to facilitate a simple mapping between visualizations and data. As a result, we were able to semantically describe standard visualizations (Bar chart, Pie chart, Scatter plot, Bubble chart, Stream graph, Line chart, Grouped Bar chart, Map, Scatter plot matrix, Parallel coordinates and Timeline) and create a system which uses this description to automatically visualize user's data. Having build it following the general information visualization principles, our vocabulary should represent a common description of statistical visualizations and thus be domain- and technology independent.
- To ensure the technological independency of our vocabulary, we applied it

to describe visualizations from D3 <sup>1</sup> visualization library, Google Charts <sup>2</sup> catalogue and from SIMILE project. <sup>3</sup> Considering the insights we gained thereby, we adapt the vocabulary by adding or removing parts used to describe visualizations. Thus, with the final version of the vocabulary, we were able to describe any visualization from the libraries mentioned above. Note, we do not claim to have approached every existing visualization libraries with our vocabulary but those which were very popular at the time, we started with the vocabulary.

Regarding to the domain-independency, with our vocabulary we can address a broader spectrum of application domains. To confirm this, we used the visualizations from the vocabulary to visualize research data (e.g., structured as RDF Data Cubes), data with cultural and educational content (provided in RDF or JSON format) and log-data (e.g., student's learning activity provided in CSV format). Nevertheless, we are aware that some domains have their specific requirements which the visualizations we currently have in our library might not address. Thus, we provide domain experts the possibility to easily add their own visualizations into the vocabulary by just following the specification as well as domain specific metadata defining a higher-level characterization of the data and customized features.

The experiences we made and the insights we gained with regard to the rule-based recommender can be summarized as follows:

- The visual mapping defines the heart of our rule-based recommender and is performed by two steps. First, the system selects the candidate visualization and the appropriate data attributes. Second, it maps the values of each attribute onto the suitable visual channel of the visualization. The process of selecting the appropriate visualizations and data attributes is guided by the visual encoding rules and guidelines. Given that the Visual Analytics Vocabulary is a collection of expert knowledge about visualization principles, it is applied in our rule-based system, together with the mapping algorithm, to automatically create appropriate visualizations. This worked well for the structured data, such as RDF Data Cubes. However, for semi-structured data we obtained  $m!$  permutations per visualization. To limit the size of possible mappings, we defined visual patterns each describing a possible set of syntactically valid mappings.
- The mapping algorithm determines a possible mapping by verifying the structural- and data type compatibility between the visual patterns and data attributes. In case of RDF Data Cubes, we automatically obtained the data type information of the attributes, of dimensions and measures respectively. However, in

---

<sup>1</sup><https://d3js.org/>

<sup>2</sup><https://developers.google.com/chart/interactive/docs/gallery>

<sup>3</sup><http://www.simile-widgets.org/>



case of semi-structured data, we had to exploit intelligent methods to detect the datatypes. For example, we used gazetteer lists to detect if an attribute value was a *country, day, city, language, geo-coordinates, etc.*

- We performed a preliminary evaluation to investigate how readily non-experts can perform complex analysis tasks using our visual recommender. Eight volunteers participated in our study. In general, the evaluation results confirmed our hypothesis that our rule-based recommender can support the ordinary user in visually exploring their data and perform complex analysis tasks. However, the number of participants was too small to perform statistical inferences. As a result, we planned our next studies in a way that collected data could be used to draw formal conclusions.

The experiences we made and the insights we gained with regard to the personalized visual recommender can be summarized as follows:

- The results we obtained for the preliminary evaluation of our rule-based recommender led to the hypothesis that the choice of visual representation involves as much user preferences and needs. To validate this hypothesis, we performed a crowd-sourced study exploiting Amazon Mechanical Turk framework. Through this study, we collected empirical evidence that a small number of visualizations are generally preferred and only a personalized approach would address user's visual preferences and needs better. Recommender systems address the personalization issue basing on knowledge about interests and previous choices (=preferences) of the users. Considering this, we investigated different recommender system approaches on their availability to be used in visual recommenders. For this investigation, we chose the most popular recommendation strategies (collaborative filtering, content-based filtering) each using a different representation method of user's preferences and needs/interest (cf., ratings, tags) to define personalized recommendations. We started with the collaborative filtering (CF), the most prominent recommendation strategy, which uses the user ratings and has been successfully applied in many domains. However, our CF-based recommender suffered from the cold start and data sparsity issues. Thus, we included the content-based approach which generates personalized visualizations based on user's current topic of interest provided in form of tags. We evaluated the performance of the two recommenders against each other performing an offline evaluation. Looking at the results, we observed that the CF-based recommender performs better than CB-based recommender when we have enough information about a user (c.f., 80% of user's preferences). However, the less information we have about the user the better was CB-based recommender. Considering these results, we decided to combine both techniques to a hybrid recommender and evaluate its performance. The results of the last evaluation implied that hybrid filtering provides better recommendations than the other two methods. These results

were not surprising, as the hybrid recommender takes the sum of both recommenders and, if one of them failed to generate recommendation, the other one would help to fill that gap and produce recommendations. In summary, the main outcomes of these experiments were (i) popular recommendation methods can be effectively applied to define personalized visual recommendations, although there is huge potential to apply more state-of-the art approaches such as matrix-factorization and (ii) there is no difference between recommending movies, books, food or visualizations since humans' individual visual preferences as much vary as their individual preferences for entertainment, food, travel etc.

- For our visual recommender we are using the user preferences provided in form of ratings or/and tags. This information is collected asking the user to explicitly rate or annotate the visualizations. Our research so far did not concentrate on whether the user is willing to provide this information for visualizations. This is a open research question for future. However, the evaluations we performed provided the evidence that the users highly benefit from rating and tagging as they produce graphical representations the users need to effectively explore their data.

The experiences we made and the insights we gained with regard to the content descriptors for visual recommender systems can be summarized as follows:

- We performed an Amazon Mechanical Turk study to collect personalized scores for visualizations. Considering the experiences researches so far made with crowd-sourced studies, we required from participants to perform preparatory tasks before rating the visualization. This should bring the participants to accurately study the visualizations and prevent a random or rash rating. As a result, we designed tasks as follows: (1) a participant was given a one-line description of a dataset originating the visualization, (2) looking at the visualization, she had to write tags (at most five), (3) provide a title, (4) rate the visualization and (5) pose a question the visualization can answer. Although providing a title and question with corresponding answer should originally bring the user to carefully process the visualizations, we figured out that they actually represent different aspects of user's interest: titles describe user's interpretation of the visualizations and using questions the user expresses which information she extracts from the visualization. Thus, we investigated how efficient they might encode user's visual preferences and needs and the content of the visualizations. The related research suggest to use information-theoretic measures to measure the encoding power of tags. Given that titles and Q&As are keywords/terms that classify information just like tags, we applied the same information-theoretic measures to evaluate the power of titles and Q&As in encoding the user and the visualizations. Looking at the results, we found

that the best user-item model combination is guaranteed when using the questions for describing the user and the answers for describing the visualizations. Beside this outcome, the evaluation has revealed that traditional information-theoretic measures can be successfully used to determine the quality of the user provided input in encoding user's visual taste and content of visualizations.

- To investigate the quality of questions and answers in generating personalized visualizations, we performed an offline evaluation applying our content-based recommender as this approach is based on the description of the user and the items. In this evaluation, we compared the recommendations generated by the question-answer based recommender with the recommendations generated by the tags-tags and title-title based recommenders. We also compared the results of the three individual recommenders (tags-tags, title-title, question-answer) against the baseline algorithm Random (RD) which simulated the behavior of our rule-based recommender. The results revealed the good quality of questions-answers in driving personalized visualizations as they produced more accurate recommendations than RD. However, the quality of this combination was not significantly better than of the tags and titles which also outperformed RD. Considering this, we extracted the following messages from this experiment: (i) the visual recommender systems are not just restricted on tags, and (ii) in such systems, the designers are free in choosing the method for preference elicitation (i.e., tagging, titling or question-answering). To investigate how these alternative content descriptors perform in combination with ratings (proven as highly powerful in encoding the user preferences in a previous study), we repeated the experiment from above this time applying our hybrid filtering. Similar to the first experiment, there were no significant differences between the results of the individual hybrids (tags-tags, title-title, question-answer) whereby all of them outperformed the baseline algorithm. Usually, one does not find visualizations isolated. The users annotate their visualizations with some text explaining what the visualization is showing or what question it is answering. Yet, the two studies from above have proven the good quality of this kind of information sources to be used for personalized visual recommender systems instead of asking user for it. As a result, a personalized visual recommender can be defined in a way that it automatically checks what the user is writing next to a visualization and use this text to define visual recommendations.

### 7.3 Future Work

This section introduces the future research direction. It is our hope that the results presented here motivate the researcher to contribute new strategies and methods which would enrich the overall visualization recommendation experience.

## 7.4 Visualization Ontology & Rule-based Visual Recommender

Some future directions for a domain- and technology independent visualization ontology have already been hinted at. The specification of the visual ontology should be adapted in a way so that the description of 3D visualizations, networks and tree becomes possible. To achieve this, the ontology should cover the following: (i) lower level of details about visualizations and not only simple facts, (ii) hierarchical relations between visual objects, (iii) relations between visual objects and graphical space and (iv) scalability of the visual objects. As a result, the rule-based recommender has to be adapted so that it can interpret and apply this information and cover 3D visualizations, trees and networks.

Basically, each visualization supports a certain presentation type, a function respectively: comparison, composition, distribution and relationship.<sup>4</sup> This is an important detail about a visualization that can be covered by the ontology. When reused by the rule-based systems, this detail can help them to create visualizations that not only reflect the properties of user's data but also user's task. For instance, if the user is interested in comparing the values of categorical data, the system can rank bar chart higher than the other charts (e.g., pie chart, scatterplot) that are candidate charts for the selected data attributes too.

## 7.5 Personalized Visual Recommender

The recent version of our visual recommender applies the user preferences to personalize the visualizations. The preferences are a collection of ratings and tags explicitly provided by the user. We choose explicit user feedback as it is recognized to be more accurate than implicit feedback in defining recommendations. However, it is a fact that the users are not always willing to rate or tag the items, although this increases their chance to obtain only those items that are closer to what they need and prefer. We did not yet evaluate the willingness of the user in providing ratings and tags about visualizations— we defined this as one of our ongoing goals. Nevertheless, considering the results of previous evaluations with recommender systems, we are planning to focus in the future on collecting implicit user preferences by observing user's actions/behavior when interacting with our visual recommender. When having only implicit feedback, we can apply matrix factorization approach which allows the incorporation of additional information sources. This recommendation technique has become very popular in recent years due to good scalability and high accuracy in defining personalized recommendations. In a nutshell, matrix factorization techniques discover the latent characteristics of the users and the items and use them to predict the unknown ratings. When we use this recommendation tech-

---

<sup>4</sup>[https://eazybi.com/blog/data\\_visualization\\_and\\_chart\\_types/](https://eazybi.com/blog/data_visualization_and_chart_types/)

nique, we will be able to define appropriate visual recommendations even if we only have implicit ratings.

For the observation of users' behavior, we plan in the future to define a personalized assisted analytics tool. This tool would apply methods and techniques to observe users' analysis steps when performing visual exploration tasks and use this information in order to infer the visual task the current user is working on (distribution, correlation, finding anomalies, finding input/output relation) and recommend the next interaction (aggregation, filtering, brushing etc.). In doing so, the system should assist the user in completing the task successfully. Machine learning algorithms such as Markov Chains, Hidden Markov Models have been proven as effective for capturing and learning from the user behavior. Thus, we consider to apply these methods on our personalized assisted analytics tool to support the user in visually exploring her data.

## 7.6 Summary

This chapter has summarized the contributions of this thesis as well as the most important insights gained. The overview is intended to set a frame for the future directions of this research that we have identified while working on this thesis.

# Chapter 8

## Appendix

This Chapter provides supplementary information to this thesis.

### 8.1 Evaluating Rule-based Recommender System

This section lists the questions and answers of our preliminary evaluation introduced in Section 3.2.3. The evaluation has been performed as a combination of the subjective assessment of usability (usability analysis) with the NASA Task Load Index. The evaluation started with background questions about the participants:

- How is your English? The evaluation was conducted in English. From 8 participants 5 declared their English skills as “Fluent” and 3 as “Okay”.
- Have you used the CODE VizRec before? We selected for this evaluation only participants with no prior experience with the CODE VizRec. Accordingly, all 8 participants answered with “no”.
- Have you used the CODE Linked Data Query Wizard before? We selected for this evaluation only participants with no prior experience with the CODE Linked Data Query Wizard. Accordingly, all 8 participants answered with “no”.
- How frequently do you use spreadsheet applications? Both CODE VizRec and the CODE Linked Data Query Wizard are mainly intended for people with prior experience with spreadsheet applications. Although this was not checked before the evaluation all participants reported to have at least some experience: 5 of 8 participants use spreadsheet applications several times a month, 2 of them once a month or less often and 1 of them every (work) day.
- How frequently do you use visualization tools? We wanted to find out, how many participants have experience with the visualization tools. 5 of 8 participants use visualization tools once a month or less often, 3 of them several times a month.

- If you use visualization tools: which tools you have experience with and what do you use them for? This question aimed to check the participants experiences with visual analysis. The answers for this question were:
  - Excel: 5 participants
  - Gnuplot: 1 participants
  - Google charts: 1 participant
  - Visio: 0 participant
  - SPSS:1 participant
  - Visual Paradigm: 0 participant
- How frequently do you look up information on the web? This question aimed to probe the level of the participants web experience. All participants indicated to look up information on the web every (work) day.
- How often do you work with Linked (Open) Data? This question was intended to find out if there were any Linked Data experts among the participants. Only 1 of 8 participants answered “Once a month or less often”, 6 of them answered “Never” and 1 of them answered “Whats Linked Open Data”.
- How frequently do you write SPARQL queries? This question was intended to find out if there were any Semantic Web experts among the participants. Only 2 of 8 participants answered “Once a month or less often”, 6 of them answered “Never”.
- Whats your age? The final background question provides information about the age range of the participants: 6 participants were between 28 and 37 and 2 of them between 18 and 27.

After the initial background questions, the participants had to complete six tasks using the CODE Visualization and CODE Query Wizard. These were the questions:

- **Task 1: Filtering in the Query Wizard** To begin the task, please click on the following link: <http://codev.knowcenter.tugraz.at/search#?endpoint=http%3A%2F%2Fcode.know-center.tugraz.at%3A8890%2Fsparql>

**Data:** This task deals with the dataset called “G 20-Countries: Co2(Tons per Person), Life Expectancy, Population; 2000-2009”.

**Task:** Please show the data set in Query Wizard. We are interested only in the countries which have a CO2 Emission over 13 Tons per persons. After that, please visualize the results.

You have 3 minutes to complete this task.

- **Task 2: Filtering in the VizRec** To begin the task, please click on the following link: <http://codev.knowcenter.tugraz.at/search#?endpoint=http%3A%2F%2Fcode.know-center.tugraz.at%3A8890%2Fsparql>

**Data:** This task deals with the same data as before, the dataset called “G 20-Countries: Co2(Tons per Person), Life Expectancy, Population; 2000-2009”.

**Task:** We are still interested only in the countries which have a CO2 emission over 13 Tons per persons. Please solve this task by using one of the suggested visualizations.

Use brushing!

Which countries have CO2 emission over 13 Tons per persons?

You have 3 minutes to complete this task.

- **Task 3: Aggregation** To begin the task, please click on the following link: <http://codev.knowcenter.tugraz.at/search#?endpoint=http%3A%2F%2Fcode.know-center.tugraz.at%3A8890%2Fsparql>

**Data:** This task deals with the same data set as before: “G 20-Countries: Co2(Tons per Person), Life Expectancy, Population; 2000-2009”.

**Task:** We are interested in the (average) life expectancy for each country.

1. Visualize the data set
2. Aggregate this dataset in such a way to obtain the (average) life expectancy for each country:
  - group by countries
  - select average as aggregation function for life expectancy
3. Please show this data in a geo chart.

Which countries have the smallest and the largest life expectancy in G20 countries?

You have 5 minutes to complete this task

- **Task 4: Aggregation - Multiple Categories** To begin the task, please click on the following link: <http://codev.knowcenter.tugraz.at/search#?endpoint=http%3A%2F%2Fcode.know-center.tugraz.at%3A8890%2Fsparql>

**Data:** This task deals with the dataset called “G 20-Countries: Co2(Tons per Person), Life Expectancy, Population; 2000-2009”.

**Task:** We want to find out how the average rate of the CO2 emission developed over the years and for different regions.

1. Visualize the data set.



2. Please show, for different regions, development of the average rate of the CO<sub>2</sub> emission over the years.
  - Use aggregation (tip: you can group over multiple categories)
  - Select the appropriate visualization (tip: change the mapping of Available Categories, if needed)

In which regions are CO<sub>2</sub> emissions rising and where are they falling?

You have 10 minutes to complete this task.

- **Task 5: Aggregating Multiple Values** To begin the task, please click on the following link: <http://codev.knowcenter.tugraz.at/search#?endpoint=http%3A%2F%2Fcode.know-center.tugraz.at%3A8890%2Fsparql>

**Data:** This task deals with the same dataset called “G 20-Countries: Co<sub>2</sub>(Tons per Person), Life Expectancy, Population; 2000-2009”.

**Task:** We want to find out whether there is a correlation between CO<sub>2</sub> emissions and life expectancy for regions.

1. Please visualize the data set.
2. Please find out whether there is a relation between average CO<sub>2</sub> emissions and the average life expectancy.

Tip: you can add multiple aggregate values!

Do you see a correlation between CO<sub>2</sub> emissions and life expectancy (for region)?

You have 12 minutes to complete this task.

- **Task 6: Brushing in Multiple Views** To begin the task, please go to prepared visualization

**Data:** This task deals with the same dataset called “G 20-Countries: Co<sub>2</sub>(Tons per Person), Life Expectancy, Population; 2000-2009”. You see visualizations showing the correlation between average CO<sub>2</sub> emissions and the average life expectancy for countries

**Task:** We want to find out on which continents the countries with

- the lowest CO<sub>2</sub> emissions and life expectancy are located?
- the highest CO<sub>2</sub> emissions and life expectancy are located?

(tip: use brushing)

You have 5 minutes to complete this task.

After each task was finished (either the participants successfully completed it or they exceeded the time limit) the participants filled out a NASA Task Load Index form which allows participants to subjectively judge the level of workload for each task. These were the questions consisted of the form:

- **Mental Demand** How mentally demanding was the task?
- **Physical Demand** How physically demanding was the task?
- **Temporal Demand** How hurried or rushed was the pace of the task?
- **Performances** How successful were you in accomplishing what you were asked to do?
- **Effort** How hard did you have to work to accomplish your level of performance?
- **Frustrations** How insecure, discouraged, irritated, stressed and annoyed were you?

The participants had also the possibility to give any comments about what was good/bad/unexpected/difficult during each task. The final page of the form consisted some additionally questions that gave the participants the opportunity to provide more qualitative feedback. These were the questions:

- What did you like about the Query Wizard?
- What did you dislike about the VizRec?
- How did you like the visual design of the VizRec?
- Do you have suggestions for improving the design of the VizRec?
- Was the VizRec easy to use and interact with?
- Do you have suggestions for improving the usability of the VizRec?
- For which tasks would you personally use the VizRec?
- Comment shortly the Query Wizard. What did you like or dislike, what would you use if for?
- If you could have solved any of the tasks with other tools of your choice, which ones would you have used?
- Was filtering (in the Query Wizard) hard or easy to use?
- Was aggregation hard or easy to use?

**Instructions**

In this survey, you will evaluate 3 charts in terms of your impression on the information they contain. For each chart you will have to perform 4 tasks:

1. write tags
2. write a title for the chart
3. rate the chart
4. write a question and give the corresponding answer

Please, take in consideration the following rules while performing this survey:

- Before You start with the survey, please **accept** the HIT.
- This survey has 5 pages. You can pass to the next page using the **next** button below right.
- Please study carefully the short **tutorial** on page 2 which gives an example about how to perform the tasks for a chart.
- The survey starts on **page 3**.
- Please consider carefully every displayed chart. In the task bar, **Description of the used Dataset**. You will find a short description of the dataset that originated the chart.
- Please execute **all 4 tasks** for each chart.
- If the chart does not show enough meaningful information, please write **not applicable** for the **tag** and **question**.
- You will receive a **message**, if you did not perform the minimum requirement for each task and can not pass to the next chart.
- There are three charts in this HIT, when you finished them all, please press **submit** to finish the survey.

1. What is your gender?

Male

Female

2. What is your age?

2. Your country?

Current page: 1 of 5 Next ->

**Figure 8.1:** Amazon Mechanical Turk Evaluation: instructions. Instructions clearly stated each step of the study. The instructions page also included demographical data.

- Was brushing hard or easy to use?
- Did you find automatic generation of visualizations helpful?
- Was choosing and configuring visualization hard or easy to achieve?

## 8.2 Amazon Mechanical Turk Study

This section illustrates the crowd-sourced experiment used for preference elicitation. A Human Intelligent Task (HIT) comprised three equal trials consisting of: studying a visualization, writing five (5) tags for it, rating the visualization using nine usability factors and, finally, write a question the visualization might answer.

The complete HIT is illustrated in Figures 8.1 to 8.5. The instructions for the HIT were given as shown in Figure 8.1. The instructions page included also demographic questions. After the introduction, the participant was introduced to the task with a tutorial presenting exemplary responses for a single trial as shown in 8.2. The tutorial was finished by pressing a button. A participant then got the three trials of the HIT as shown in Figure 8.3, Figure 8.4 and Figure 8.5. A warning was shown when a participant failed to write tags. It was not possible to finish a trial without rating the visualization in every dimension.

**Tutorial**

- You see on this page an exemplary chart, which is generated for a dataset about **CO2 emissions**.
- On the right side of the chart you see a box with **tags**, describing the information the user can provide by considering this chart.
- Below the chart, you see how the **rating** of this chart can look like.
- Below the rating box, you see which **questions** user can ask and **answer** using this chart.

After studying the tutorial please click on **next** button to start with survey.

Please write at least 2 (max. 5) tags describing this chart. Hint: Think about what search terms would you apply to find this chart!

Tag 1

Tag 2

Tag 3

Tag 4

Tag 5

Please write a title for this chart

This is a possible title for the displayed chart

Please rate the chart above between 1 (means not applicable) and 7 (means very applicable)

Easy to understand	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input checked="" type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Cluttered	<input type="radio"/> 1 <input type="radio"/> 2 <input checked="" type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Useful	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Boring	<input type="radio"/> 1 <input checked="" type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Organized, clear	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Exciting, interesting	<input type="radio"/> 1 <input type="radio"/> 2 <input checked="" type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Effective	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input checked="" type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Satisfying	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input checked="" type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Confusing	<input type="radio"/> 1 <input checked="" type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7

Please write 1 question, that can be answered by looking the chart

Question 1

Answer 1

This is a question which can be answered by looking the displayed chart

Current page: 2 of 5

Next →

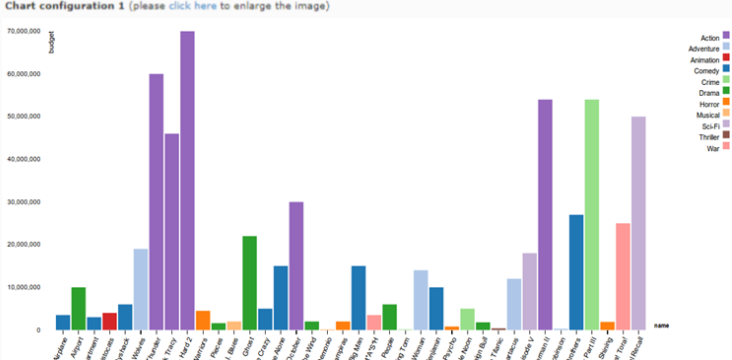
**Figure 8.2:** Amazon Mechanical Turk Evaluation: tutorial. The tutorial showed an exemplary task, with sample responses for tags, title, ratings and question&answer.

**Description of the Dataset**

The chart below is about **movies**.

---

**Chart configuration 1** (please [click here](#) to enlarge the image)



**Please write at least 2 (max. 5) tags** describing this chart. Hint: Think about what search terms would you apply to find this chart!

---

**Please write a title for this chart**

---

**Please rate the chart above between 1 (means not applicable) and 7 (means very applicable)**

Easy to understand	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Cluttered	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Useful	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Boring	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Organized, clear	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Exciting, interesting	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Effective	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Satisfying	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Confusing	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7

---

**Please write 1 question, that can be answered by looking the chart**

---

Current page: 3 of 5
[Next](#) →

**Figure 8.3:** Amazon Mechanical Turk Evaluation: first visualization to evaluate.

## 8.2.1 Demographics

47 unique workers participated in our Amazon Mechanical Turk study. Some demographic details are listed below:

*Gender:* male = 30, female = 17.

**Description of the Dataset**

The chart below is about movies.

---

**Chart configuration 1** (please [click here](#) to enlarge the image)

Please write at least 2 (max. 5) tags describing this chart. Hint: Think about what search terms would you apply to find this chart!

Tag 1

Tag 2

Tag 3

Tag 4

Tag 5

---

**Please write a title for this chart**

---

**Please rate the chart above between 1 (means not applicable) and 7 (means very applicable)**

Easy to understand	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Cluttered	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Useful	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Boring	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Organized, clear	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Exciting, interesting	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Effective	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Satisfying	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Confusing	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7

---

**Please write 1 question, that can be answered by looking the chart**

Question 1

Answer 1

---

Current page: 4 of 5
Next →

**Figure 8.4:** Amazon Mechanical Turk Evaluation: second visualization to evaluate.

*Age:* below 25 = 5, between 25 and 34 = 26, between 35 and 44 = 13 and above 44 = 3.

*Country:* USA = 42, India: 4, Iceland = 1.

Description of the Dataset

The chart below is about movies.

---

Chart configuration 1 (please click here to enlarge the image)

Legend:

- France
- Italy
- Japan
- South Africa
- Spain
- United Kingdom
- United States of America

Please write at least 2 (max. 5) tags describing this chart. Hint: Think about what search terms would you apply to find this chart!

Tag 1

Tag 2

Tag 3

Tag 4

Tag 5

---

Please write a title for this chart

---

Please rate the chart above between 1 (means not applicable) and 7 (means very applicable)

Easy to understand	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Cluttered	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Useful	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Boring	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Organized, clear	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Exciting, interesting	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Effective	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Satisfying	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Confusing	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7

---

Please write 1 question, that can be answered by looking the chart

Question 1

Answer 1

---

Current page: 5 of 5

**Figure 8.5:** Amazon Mechanical Turk Evaluation: third visualization to evaluate.

## 8.3 Visual Analytics Vocabulary

This section presents the Visual Analytics Vocabulary. The Visual Analytics (VA) Vocabulary is an explicit conceptualization, a model, that describes the visualizations in pragmatic, simple facts that guides the sensible mapping from data to visual representations. The vocabulary has been inspired to a large extent by Bertin's

---

theory on information visualization and Mackinlay's formal language to generate graphical presentations (see Section 2.1.2). The whole vocabulary is shown in Listing 8.1.

## 8.4 Visualization Ontology

This section presents the Visualization Ontology which originates from the Visual Analytics Vocabulary and provides semantics to the visualizations. Listing 8.2 illustrates the Visualization Ontology which describes eleven individual visualizations: bar chart, pie chart, scatter plot, bubble chart, stream graph, line chart, grouped bar chart, map, scatter plot matrix, parallel coordinates and timeline.



## Listing 8.1: Visual Analytics Vocabulary.

```

# Copyright (C) 2014
# "Kompetenzzentrum_fuer_wissensbasierte_Anwendungen_Forschungs-_und_EntwicklungsgmbH"
# (Know-Center), Graz, Austria, office@know-center.at.
#
# Licensees holding valid Know-Center Commercial licenses may use this file in
# accordance with the Know-Center Commercial License Agreement provided with
# the Software or, alternatively, in accordance with the terms contained in
# a written agreement between Licensees and Know-Center.
#
# This program is free software: you can redistribute it and/or modify
# it under the terms of the GNU Affero General Public License as
# published by the Free Software Foundation, either version 3 of the
# License or (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
# GNU Affero General Public License for more details.
#
# You should have received a copy of the GNU Affero General Public License
# along with this program. If not, see <http://www.gnu.org/licenses/>.

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xs: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix sio: <http://semanticscience.org/ontology/sio.owl> .
@prefix va: <http://code-research.eu/ontology/visual-analytics#> .
@prefix qb: <http://purl.org/linked-data/cube#> .
@prefix vann: <http://purl.org/vocab/vann/> .

<http://code-research.eu/ontology/visual-analytics>
  a owl:Ontology;
  owl:versionInfo "0.1.20121122";
  rdfs:label "The_Visual_Analytics_Vocabulary";
  rdfs:comment "This_vocabulary_allows_the_semantic_description_of_visual_analytics_applications.It_is_based_
    on_the_RDF_Data_Cube_Vocabulary_and_the_Semanticscience_Integrated_Ontology.";
  dcterms:created "2012-10-31"^^xs:date;
  dcterms:modified "2014-03-24"^^xs:date;
  vann:preferredNamespacePrefix "va";
  vann:preferredNamespaceUri "http://code-research.eu/ontology/visual-analytics#";
  dc:creator "Belgin_Mutlu,_Patrick_Hoefler";
  dc:contributor "Michael_Granitzer";
  .

# === Classes =====

# --- General Classes -----

va:Chart a rdfs:Class, owl:Class;
  rdfs:label "Chart"@en;
  rdfs:comment "Is_the_graphical_representation_of_numerical_or_qualitative_data."@en;
  owl:equivalentClass sio:SIO_000904;
  va:hasChartName va:ChartName;
  va:supportedDimension va:SupportedDimension;
  va:hasVisualChannel va:VisualChannel;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
  .

va:VisualChannel a rdfs:Class, owl:Class;
  rdfs:label "Visual_channel"@en;
  rdfs:comment "Represents_a_visual_dimension_of_a_chart."@en;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
  va:hasDataType va:DataType;
  va:hasOccurrence va:Occurrence;
  va:hasPersistence va:Persistence;
  .

```

```
va:ChartName a rdfs:Class, owl:Class;
  rdfs:label "Chart_Name"@en;
  rdfs:comment "Represents the name of a chart."@en;
  rdfs:subClassOf va:Chart;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:Axis a rdfs:Class, owl:Class;
  rdfs:label "Axis"@en;
  rdfs:comment "Is a line segment that is part of a chart in which the position along the line
  corresponds to a numeric or categorical value."@en;
  rdfs:subClassOf va:VisualChannel;
  owl:equivalentClass sio:SIO_000450;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:Color a rdfs:Class, owl:Class;
  rdfs:label "Color"@en;
  rdfs:comment "Represents a visual dimension of a chart."@en;
  rdfs:subClassOf va:VisualChannel;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:Size a rdfs:Class, owl:Class;
  rdfs:label "Size"@en;
  rdfs:comment "Represents a visual dimension of a chart."@en;
  rdfs:subClassOf va:VisualChannel;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:Symbol a rdfs:Class, owl:Class;
  rdfs:label "Symbol"@en;
  rdfs:comment "Represents a visual dimension of a chart."@en;
  rdfs:subClassOf va:VisualChannel;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:DataType a rdfs:Class, owl:Class;
  rdfs:label "Data_type"@en;
  rdfs:comment "Data_type identifies a type of data represented in charts visual channel."@en;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:Persistence a rdfs:Class, owl:Class;
  rdfs:label "Persistence"@en;
  rdfs:comment "Persistence denotes whether a visual channel is permanently present in the chart and
  must be specified or it might be defined if needed."@en;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:Occurrence a rdfs:Class, owl:Class;
  rdfs:label "Occurrence"@en;
  rdfs:comment "Occurrence denotes whether a visual channel can be set only one times or multiple
  times."@en;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:Mandatory a rdfs:Class, owl:Class;
  rdfs:label "Mandatory_Persistence"@en;
  rdfs:comment "Denotes that a visual channel must be specified."@en;
  rdfs:subClassOf va:Persistence;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:Optional a rdfs:Class, owl:Class;
  rdfs:label "Optional_Persistence"@en;
  rdfs:comment "Denotes that a visual channel must be specified."@en;
  rdfs:subClassOf va:Persistence;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.
```

```

va:One a rdfs:Class, owl:Class;
  rdfs:label "occurrence_one"@en;
  rdfs:comment "Denotes that a visual channel can only set at one times."@en;
  rdfs:subClassOf va:Occurrence;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:Multiplicity a rdfs:Class, owl:Class;
  rdfs:label "occurrence_multiplicity"@en;
  rdfs:comment "Denotes that a visual channel can set at multiple times."@en;
  rdfs:subClassOf va:Occurrence;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:SupportedDimension a rdfs:Class, owl:Class;
  rdfs:label "Supported_dimension"@en;
  rdfs:comment "Denotes how many dimensions can be visualize by a chart."@en;
  rdfs:subClassOf va:Chart;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

# --- Mapping-Specific Classes -----
va:Mapping a rdfs:Class, owl:Class;
  rdfs:label "Mapping"@en;
  rdfs:comment "This class contains all mappings created for an RDF Data Cube."@en;
  va:hasComponentMapping va:ComponentMapping;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

va:ComponentMapping a rdfs:Class, owl:Class;
  rdfs:label "Component_mapping"@en;
  rdfs:comment "Represents a mapping for a particular chart."@en;
  va:representsChart va:MappedChartName;
  va:hasComponentMapping va:ComponentMapping;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

# --- Dashboard-Specific Classes -----
va:Dashboard a rdfs:Class, owl:Class;
  rdfs:label "Dashboard"@en;
  rdfs:comment "Represents a collection of one or more charts that visualize the same Data Cube."@en;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

# === Properties =====
# --- General Properties -----
va:unit a rdf:Property, owl:ObjectProperty;
  rdfs:label "Unit"@en;
  rdfs:comment "Defines the unit measure of a visual channel."@en;
  rdfs:domain va:VisualChannel;
  rdfs:range rdfs:Resource;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

owl:Restriction owl:onProperty va:unit;
  owl:minCardinality 1;
  owl:maxCardinality 1;
.

```

```

va:hasDataType a rdf:Property, owl:ObjectProperty;
  rdfs:label "Has_data_type"@en;
  rdfs:comment "Defines which data type a visual channel has."@en;
  rdfs:domain va:VisualChannel;
  rdfs:range va:DataType;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

owl:Restriction owl:onProperty va:hasDataType ;
  owl:minCardinality 1;
.

va:hasPersistence a rdf:Property, owl:ObjectProperty;
  rdfs:label "Has_persistence"@en;
  rdfs:comment "Denotes which visual channel of a chart must be specified or might be defined if needed."@en;
  rdfs:domain va:VisualChannel;
  rdfs:range va:Persistence;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

owl:Restriction owl:onProperty va:hasPersistence;
  owl:minCardinality 1;
  owl:maxCardinality 1;
.

va:hasOccurrence a rdf:Property, owl:ObjectProperty;
  rdfs:label "Has_occurrence"@en;
  rdfs:comment "Denotes whether a visual channel can be set only at one times or multiple times."@en
  ;
  rdfs:domain va:VisualChannel;
  rdfs:range va:occurrence;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

owl:Restriction owl:onProperty va:hasOccurrence;
  owl:minCardinality 1;
  owl:maxCardinality 1;
.

va:hasVisualChannel a rdf:Property, owl:ObjectProperty;
  rdfs:label "Has_visual_channel"@en;
  rdfs:comment "Represents which visual channel a chart has."@en;
  rdfs:domain va:Chart;
  rdfs:range va:VisualChannel;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

owl:Restriction owl:onProperty va:hasVisualChannel;
  owl:minCardinality 1;
.

va:hasChartName a rdf:Property, owl:ObjectProperty;
  rdfs:label "Has_chart_label"@en;
  rdfs:comment "Represents which name a chart has."@en;
  rdfs:domain va:Chart;
  rdfs:range va:ChartName;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

owl:Restriction owl:onProperty va:hasChartName ;
  owl:minCardinality 1;
  owl:maxCardinality 1;
.

va:cubeDimensionNominal a qb:DimensionProperty, rdf:Property, owl:ObjectProperty;
  rdfs:label "Nominal_cube_dimension";
  rdfs:comment "This class can be used as a fallback for nominal dimensions. It should be avoided in favor of more specific dimension properties.";
  rdfs:domain qb:Observation;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
.

```

```

# --- Mapping-Specific Properties -----

va:getMapping a rdf:Property, owl:ObjectProperty;
  rdfs:label "Get_Mapping"@en;
  rdfs:comment "Displays_the_mappings_for_a_RDF_Data_Cube."@en;
  rdfs:domain va:RDFDataCube;
  rdfs:range va:Mapping;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
  .

owl:Restriction owl:onProperty va:getMapping;
  owl:minCardinality 1;
  .

va:hasComponentMapping a rdf:Property, owl:ObjectProperty;
  rdfs:label "Has_components_mapping"@en;
  rdfs:comment "Displays_a_particular_mapping_of_a_RDF_Data_Cube."@en;
  rdfs:domain va:Mapping;
  rdfs:range va:ComponentMapping;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
  .

owl:Restriction owl:onProperty va:hasComponentsMapping;
  owl:minCardinality 1;
  .

va:representsCubeProperty a rdf:Property, owl:ObjectProperty;
  rdfs:label "Represents_chart"@en;
  rdfs:comment "Represents_the_component_(dimension,measure)_of_a_RDF_Data_Cube,_which_has_been_
  mapped."@en;
  rdfs:domain va:Mapping;
  rdfs:range qb:DimensionProperty;
  rdfs:range qb:MeasureProperty;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
  .

owl:Restriction owl:onProperty va:representsCubeProperty;
  owl:minCardinality 1;
  owl:maxCardinality 1;
  .

# --- Dashboard-Specific Properties -----

va:visualizes a rdf:Property, owl:ObjectProperty;
  rdfs:label "visualizes"@en;
  rdfs:comment "Links_a_Visual_Analytics_Dashboard_with_a_Data_Cube_or_Data_Slice"@en;
  rdfs:domain va:Dashboard;
  rdfs:range qb:DataSet, qb:Slice;
  rdfs:isDefinedBy <http://code-research.eu/ontology/visual-analytics>;
  .

owl:Restriction owl:onProperty va:visualizes;
  owl:minCardinality 1;
  owl:maxCardinality 1;
  .

```

## Listing 8.2: Visualization Ontology.

```

<?xml version="1.0"?>
<rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:vo="http://eexcess.eu/
  visualisation-ontology#" xmlns:va="http://code-research.eu/ontology/visual-analytics#" xmlns:rdf="
  "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <va:BarChart rdf:about="http://eexcess.eu/visualisation-ontologyBarChart">
    <rdfs:label>barchart</rdfs:label>
    <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-analyticsbarchart" />
    <vo:hasVisualChannel>
      <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyBarChartXAxis">
        <rdfs:label>x-Axis</rdfs:label>
        <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
        <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#location" />
        <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#date" />
        <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
          Mandatory" />
        <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
          " />
      </vo:Axis>
    </vo:hasVisualChannel>
    <vo:hasVisualChannel>
      <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyBarChartYAxis">
        <rdfs:label>y-Axis</rdfs:label>
        <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
        <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
        <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
          Mandatory" />
        <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
          " />
      </vo:Axis>
    </vo:hasVisualChannel>
  </va:BarChart>

  <va:PieChart rdf:about="http://eexcess.eu/visualisation-ontologyPieChart">
    <rdfs:label>piechart</rdfs:label>
    <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-analyticspiechart" />
    <vo:hasVisualChannel>
      <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologypieChartXAxis">
        <rdfs:label>x-Axis</rdfs:label>
        <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
        <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
          Mandatory" />
        <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
          " />
      </vo:Axis>
    </vo:hasVisualChannel>
    <vo:hasVisualChannel>
      <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologypieChartYAxis">
        <rdfs:label>y-Axis</rdfs:label>
        <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
        <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
        <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
          Mandatory" />
        <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
          " />
      </vo:Axis>
    </vo:hasVisualChannel>
  </va:PieChart>

```

```

<va:ScatterPlot rdf:about="http://eexcess.eu/visualisation-ontologyScatterplot" >
  <rdfs:label>scatterplot</rdfs:label>
  <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-analyticsscatterplot" /
  >
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyscatterplotXAxis" >
      <rdfs:label>x-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
      <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
      Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
      " />
    </vo:Axis>
  </vo:hasVisualChannel>
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyscatterplotYAxis" >
      <rdfs:label>y-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
      <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
      Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
      " />
    </vo:Axis>
  </vo:hasVisualChannel>
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyscatterplotColor" >
      <rdfs:label>Color</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
      Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
      " />
    </vo:Axis>
  </vo:hasVisualChannel>
</va:ScatterPlot>

<va:BubbleChart rdf:about="http://eexcess.eu/visualisation-ontologyBubblechart" >
  <rdfs:label>bubblechart</rdfs:label>
  <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-analyticsbubblechart"
  />
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologybubblechartXAxis" >
      <rdfs:label>x-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
      <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
      Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
      " />
    </vo:Axis>
  </vo:hasVisualChannel>

```

```

<vo:hasVisualChannel>
  <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologybubblechartYAxis">
    <rdfs:label>y-Axis</rdfs:label>
    <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
    <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
    <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
    <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
  </vo:Axis>
</vo:hasVisualChannel>
<vo:hasVisualChannel>
  <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologybubblechartSize">
    <rdfs:label>Size</rdfs:label>
    <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
    <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
    <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
    <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
  </vo:Axis>
</vo:hasVisualChannel>
<vo:hasVisualChannel>
  <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologybubblechartColor">
    <rdfs:label>Color</rdfs:label>
    <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
    <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
    <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
  </vo:Axis>
</vo:hasVisualChannel>
</va:BubbleChart>

<va:Streamgraph rdf:about="http://eexcess.eu/visualisation-ontologyStreamgraph">
  <rdfs:label>streamgraph</rdfs:label>
  <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-analyticsstreamgraph" />
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologystreamgraphXAxis">
      <rdfs:label>x-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#date" />
      <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#date" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
    </vo:Axis>
  </vo:hasVisualChannel>

```



```

<vo:hasVisualChannel>
  <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologystreamgraphYAxis">
    <rdfs:label>y-Axis</rdfs:label>
    <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
    <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
    <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
    <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
  </vo:Axis>
</vo:hasVisualChannel>
<vo:hasVisualChannel>
  <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologystreamgraphColor">
    <rdfs:label>Color</rdfs:label>
    <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
    <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
    <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
  </vo:Axis>
</vo:hasVisualChannel>
</va:Streamgraph>

<va:LineChart rdf:about="http://eexcess.eu/visualisation-ontologyLinechart">
  <rdfs:label>linechart</rdfs:label>
  <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-analyticslinechart" />
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyLineChartXAxis">
      <rdfs:label>x-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#date" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
    </vo:Axis>
  </vo:hasVisualChannel>
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyLineChartYAxis">
      <rdfs:label>y-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
      <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
    </vo:Axis>
  </vo:hasVisualChannel>
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://code-research.eu/ontology/visual-analyticslineChartLines">
      <rdfs:label>Lines</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#location" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
    </vo:Axis>
  </vo:hasVisualChannel>
</va:LineChart>

```

```

<va:Groupedbarchart rdf:about="http://eexcess.eu/visualisation-ontologyGroupedbarchart">
  <rdfs:label>groupedbarchart</rdfs:label>
  <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-
analyticsgroupedbarchart" />
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologygroupedbarchartXAxis">
      <rdfs:label>x-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
      <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#date" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
" />
    </vo:Axis>
  </vo:hasVisualChannel>
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologygroupedbarchartYAxis">
      <rdfs:label>y-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
      <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
" />
    </vo:Axis>
  </vo:hasVisualChannel>
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologygroupedbarchartBar">
      <rdfs:label>Bar</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
" />
    </vo:Axis>
  </vo:hasVisualChannel>
</va:Groupedbarchart>

<va:Map rdf:about="http://eexcess.eu/visualisation-ontologyMap">
  <rdfs:label>map</rdfs:label>
  <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-analyticsgroupedmap"
/>
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologymapCountry">
      <rdfs:label>country</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#location" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
" />
    </vo:Axis>
  </vo:hasVisualChannel>

```

```

    <vo:hasVisualChannel>
      <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologymapNuance" >
        <rdfs:label>nuance</rdfs:label>
        <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
        <vo:hasDataType rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
        <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
          Optional" />
        <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One
          " />
      </vo:Axis>
    </vo:hasVisualChannel>
  </va:Map>

<va:Scatterplotmatrix rdf:about="http://eexcess.eu/visualisation-ontologyScatterplotmatrix" >
  <rdfs:label>scatterplotmatrix</rdfs:label>
  <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-analyticsscatterplotmatrix"
  />
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyScatterplotmatrixAxis" >
      <rdfs:label>Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#integer" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
        Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#
        Multiplicity" />
    </vo:Axis>
  </vo:hasVisualChannel>
</va:Scatterplotmatrix>

<va:ParallelCoordinates rdf:about="http://eexcess.eu/visualisation-ontologyParallelCoordinates" >
  <rdfs:label>parallelcoordinates</rdfs:label>
  <va:hasChartName rdf:resource="http://code-research.eu/ontology/visual-
    analyticsparallelcoordinates" />
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyParallelCoordinatesXAxis" >
      <rdfs:label>x-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#integer" />
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#date" />
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#location" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#
        Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#
        Multiplicity" />
    </vo:Axis>
  </vo:hasVisualChannel>

```

```

<vo:hasVisualChannel>
  <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyParallelCoordinatesColor">
    <rdfs:label>Color</rdfs:label>
    <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
    <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#date" />
    <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#location" />
    <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
    <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#location" />
    <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
  >
</vo:Axis>
</vo:hasVisualChannel>
</va:ParallelCoordinates>

<va:TimeLine rdf:about="http://eexcess.eu/visualisation-ontologyTimeline">
  <rdfs:label>timeline</rdfs:label>
  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyTimelineXAxis">
      <rdfs:label>x-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#date" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
    </vo:Axis>
  </vo:hasVisualChannel>

  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyTimelineYAxis">
      <rdfs:label>y-Axis</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#number" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
    </vo:Axis>
  </vo:hasVisualChannel>

  <vo:hasVisualChannel>
    <vo:Axis rdf:about="http://eexcess.eu/visualisation-ontologyTimelineColor">
      <rdfs:label>color</rdfs:label>
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#string" />
      <vo:hasDataType rdf:resource="http://eexcess.eu/visualisation-ontology#location" />
      <va:hasPersistence rdf:resource="http://code-research.eu/ontology/visual-analytics#Mandatory" />
      <va:hasOccurrence rdf:resource="http://code-research.eu/ontology/visual-analytics#One" />
    </vo:Axis>
  </vo:hasVisualChannel>
</va:TimeLine>
</rdf:RDF>

```



# Bibliography

- [AB09] Jae-Wook Ahn and Peter Brusilovsky. Adaptive Visualization of Search Results: Bringing User Models to Visual Analytics. *Information Visualization*, 8(3):167–179, 2009.
- [AH11] Dean. Allemang and James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Elsevier Science, 2011.
- [BCC10] Alejandro Bellogín, Iván Cantador, and Pablo Castells. A study of heterogeneity in recommendations for a social music service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pages 1–8. ACM, 2010.
- [Ber83] Jacques Bertin. *Semiology of graphics*. University of Wisconsin Press, 1983.
- [BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.
- [Bur02] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.
- [BVB<sup>+</sup>13] Michelle A. Borkin, Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013.
- [BVdB09] Toine Bogers and Antal Van den Bosch. Collaborative and content-based filtering for item recommendation on social bookmarking websites. In *Proceedings of the ACM Recommender Systems workshop on Recommender Systems and the Social Web*, volume 9, pages 9–16, 2009.
- [Car03] M. S. T. Carpendale. Considering visual variables as a basis for information visualisation. Technical report, University of Calgary, Calgary, AB, 2003.

- [CDC<sup>+</sup>07] Mike Cammarano, Xin Dong, Bryan Chan, Jeff Klingner, Justin Talbot, Alon Halevy, and Pat Hanrahan. Visualization of heterogeneous data. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1200–1207, 2007.
- [CJ10] Min Chen and Heike Jnicke. An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1206–1215, Nov 2010.
- [CM85] W.S. Cleveland and R. McGill. Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science*, 229(4716):828, 1985.
- [CM08] Ed H. Chi and Todd Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, pages 81–88. ACM, 2008.
- [CMS99] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.
- [CRT13] Richard Cyganiak, Dave Reynolds, and Jeni Tennison. The rdf data cube vocabulary. *W3C Recommendation (January 2014)*, 2013.
- [CWT<sup>+</sup>08] Bryan Chan, Leslie Wu, Justin Talbot, Mike Cammarano, and Pat Hanrahan. Vispedia\*: Interactive visual exploration of wikipedia data via search-based integration. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1213–1220, 2008.
- [DBDH05] D. J. Duke, K. W. Brodlie, D. A. Duce, and I. Herman. Do you see what i mean? [data visualization]. *IEEE Computer Graphics and Applications*, 25(3):6–9, May 2005.
- [DFVR10] Michel Dumontier, Leo Ferres, and Natalia Villanueva-Rosales. Modeling and querying graphical representations of statistical data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2):241 – 254, 2010. Bridging the GapData Mining and Social Network Analysis for Integrating Semantic Web and Web 2.0 The Future of Knowledge Dissemination: The Elsevier Grand Challenge for the Life Sciences.
- [EB12] Micheline Elias and Anastasia Bezerianos. Annotating bi visualization dashboards: Needs & challenges. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1641–1650, New York, NY, USA, 2012. ACM.

- [FPL<sup>+</sup>06] Leo Ferres, Avi Parush, Zhihong H Li, Yandu Oppacher, and Gitte Lindgaard. Representing and querying line graphs in natural language: The iGraph system. *Smart Graphics, Proceedings*, 4073:248–253, 2006.
- [FSCG07] Umer Farooq, Yang Song, John M. Carroll, and C. Lee Giles. Social bookmarking for scholarly digital libraries. *Internet Computing, IEEE*, 11(6):29–35, November 2007.
- [Gra90] Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag New York, Inc., 1990.
- [GSGC08a] O. Gilson, N. Silva, P. W. Grant, and M. Chen. From web data to visualization via ontology mapping. In *Proceedings of the 10th Joint Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'08, pages 959–966, Chichester, UK, 2008. The Eurographs Association &#38; John Wiley &#38; Sons, Ltd.
- [GSGC08b] O Gilson, N Silva, P W Grant, and Min Chen. From Web Data to Visualization via Ontology Mapping. *Computer Graphics Forum*, 27(3):959–966, 2008.
- [GSK<sup>+</sup>99] Nathaniel Good, J. Ben Schafer, Joseph A. Konstan, Al Borchers, Badrul Sarwar, Jon Herlocker, and John Riedl. Combining collaborative filtering with personal agents for better recommendations. In *Proc. of the 16th Nat. Conf. on Artif. Intell. and the 11th Innov. Applic. of Artif. Intell. Conf. Innov. Applic. of Artif. Intell.*, AAAI '99/IAAI '99, pages 439–446, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence.
- [GVKD10] Sten Govaerts, Katrien Verbert, Joris Klerkx, and Erik Duval. *Visualizing Activities for Self-reflection and Awareness*, pages 91–100. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
- [GW09] David Gotz and Zhen Wen. Behavior-Driven Visualization Recommendation. *CHI '09 Proceedings of the 14th international conference on Intelligent user interfaces*, pages 315–324, 2009.
- [HB10a] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *ACM Human Factors in Computing Systems (CHI)*, pages 203–212, 2010.
- [HB10b] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *ACM Human Factors in Computing Systems (CHI)*, pages 203–212, 2010.



- [HGVS14] Patrick Hoefler, Michael Granitzer, Eduardo E. Veas, and Christin Seifert. Linked data query wizard: A novel interface for accessing SPARQL endpoints. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014.*, 2014.
- [HKTR04] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, January 2004.
- [HM01] I Herman and MS Marshall. GraphXMLan XML-based graph description format. *Graph Drawing*, pages 52–62, 2001.
- [JEKJ07] Jeroen Janssen, Gijbert Erkens, Gellof Kanselaar, and Jos Jaspers. Visualization of participation: Does it contribute to successful computer-supported collaborative learning? *Comput. Educ.*, 49(4):1037–1065, December 2007.
- [JM04] Jiwoon Jeon and R. Manmatha. *Using Maximum Entropy for Automatic Image Annotation*, pages 24–32. Springer Berlin Heidelberg, 2004.
- [JZFF10] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender Systems*. Cambridge University Press, 2010. Cambridge Books Online.
- [KC16] Sneha Khatwani and M. B. Chandak. Building personalized and non personalized recommendation systems. In *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pages 623–628, Sept 2016.
- [KCS08] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08*, pages 453–456, New York, NY, USA, 2008. ACM.
- [KG13] Wahiba Ben Abdesslem Karaa and Nidhal Griba. Information retrieval with porter stemmer: A new version for english. In *Advances in Computational Science, Engineering and Information Technology*, volume 225 of *Advances in Intelligent Systems and Computing*, pages 243–254. Springer International Publishing, 2013.
- [KHPA12] Alicia Key, Bill Howe, Daniel Perry, and Cecilia Aragon. VizDeck: Self-organizing dashboards for visual analytics. *Proceedings of the 2012 international conference on Management of Data - SIGMOD '12*, page 681, 2012.

- [KK13] Stefan Klampff and Roman Kern. *An Unsupervised Machine Learning Approach to Body Text and Table of Contents Extraction from Digital Scientific Articles*, pages 144–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
- [KKMN10] Ville Karavirta, Ari Korhonen, Lauri Malmi, and Thomas Naps. A comprehensive taxonomy of algorithm animation languages. *Journal of Visual Languages and Computing*, 21(1):1–22, 2010.
- [KMM<sup>+</sup>16] W. A. D. Kanchana, G. D. L. Madushanka, H. P. Maduranga, M. D. M. Udayanga, D. A. Meedeniya, and G. I. U. S. Perera. Context aware recommendation for data visualization. In *Proceedings of the 2Nd International Conference on Communication and Information Processing, ICCIP '16*, pages 22–26, New York, NY, USA, 2016. ACM.
- [KO17] Pawandeep Kaur and Michael Owonibi. A Review on Visualization Recommendation Strategies. In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, (February), 2017.
- [KRK<sup>+</sup>15] Evangelos Kalampokis, Bill Roberts, Areti Karamanou, Efthimios Tambouris, and Konstantinos A Tarabanis. Challenges on developing tools for exploiting linked open data cubes. In *Proceedings of the 3rd International Workshop on Semantic Statistics co-located with 14th International Semantic Web Conference (ISWC 2015)*, 2015.
- [KS15a] Nikolaos Konstantinou and Dimitrios-Emmanuel Spanos. *Introduction: Linked Data and the Semantic Web*, pages 1–16. Springer International Publishing, Cham, 2015.
- [KS15b] Nikolaos Konstantinou and Dimitrios-Emmanuel Spanos. *Technical Background*, pages 17–49. Springer International Publishing, Cham, 2015.
- [KVWW07] Serguei Krivov, Ferdinando Villa, Richard Williams, and Xindong Wu. *On Visualization of OWL Ontologies*, pages 205–221. Springer US, Boston, MA, 2007.
- [LdGS11] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*, pages 73–105. Springer US, 2011.
- [LFK<sup>+</sup>13] Sharon Lin, Julie Fortuna, Chinmay Kulkarni, Maureen Stone, and Jeffrey Heer. Selecting semantically-resonant colors for data visualization. In *Proceedings of the 15th Eurographics Conference on Visualization, EuroVis '13*, pages 401–410, 2013.

- [LTBH15] Yi-Ling Lin, Christoph Trattner, Peter Brusilovsky, and Daqing He. The impact of image descriptions on user tagging behavior: A study of the nature and functionality of crowdsourced tags. *Journal of the Association for Information Science and Technology*, 66(9):1785–1798, 2015.
- [LTP<sup>+</sup>15] Santiago Larrain, Christoph Trattner, Denis Parra, Eduardo Graells-Garrido, and Kjetil Nørnvåg. Good times bad times: A study on recency effects in collaborative filtering for social tagging. In *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys '15*, pages 269–272, New York, NY, USA, 2015. ACM.
- [LXTM11] M Leida, D Xiaofeng, P Taylor, and B Majeed. Toward automatic generation of SPARQL result set visualizations: A use case in service monitoring. *e-Business (ICE-B), 2011 Proceedings of the International Conference on*, (January):1–6, 2011.
- [LZMT15] Jared Lorince, Sam Zorowitz, Jaimie Murdock, and Peter M. Todd. The wisdom of the few? ”supertaggers” in collaborative tagging systems. *Journal of Web Science*, 1(1):16–32, 2015.
- [Mac86] Jock Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, April 1986.
- [Maj08] Zoran Majkić. Intensional semantics for rdf data structures. In *Proceedings of the 2008 International Symposium on Database Engineering & Applications, IDEAS '08*, pages 69–77, New York, NY, USA, 2008. ACM.
- [MC08] Qiaozhu Mei and Kenneth Church. Entropy of search logs: How hard is search? with personalization? with backoff? In *Proceedings of the International Conference on Web Search and Data Mining*, pages 45–54. ACM, 2008.
- [MHS07] Jock Mackinlay, Pat Hanrahan, and Chris Stolte. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, Nov 2007.
- [MHS<sup>+</sup>13] Belgin Mutlu, Patrick Höfler, Vedran Sabol, Gerwald Tschinkel, and Michael Granitzer. Automated visualization support for linked research data. In *Proceedings of the I-SEMANTICS 2013 Posters & Demonstrations Track, Graz, Austria, September 4-6, 2013*, pages 40–44, 2013.
- [MHT<sup>+</sup>14] Belgin Mutlu, Patrick Höfler, Gerwald Tschinkel, Eduardo E. Veas, Vedran Sabol, Florian Stegmaier, and Michael Granitzer. Suggesting

- visualisations for published data. In *Proceedings of the 5th International Conference on Information Visualization Theory and Applications, IVAPP 2014, Lisbon, Portugal, 5-8 January, 2014.*, pages 267–275, 2014.
- [MMVM0] Pere Millan-Martinez and Pedro Valero-Mora. Automating statistical diagrammatic representations with data characterization. *Information Visualization*, 0(0):1473871617715326, 0.
- [MS15] Belgin Mutlu and Vedran Sabol. Visual analysis of scientific content. *IEEE Special Technical Community Social Networking*, 3(1), 2015.
- [MTD<sup>+</sup>07] Joerg Meyer, Jim Thomas, Stephan Diehl, Brian Fisher, D Keim, D Laidlaw, Silvia Miksch, Klaus Mueller, William Ribarsky, Bernhard Preim, and Anders Ynnerman. From Visualization to Visually Enabled Reasoning. *Scientific Visualization Advanced Concepts*, 1:227–245, 2007.
- [Mun14] Tamara Munzner. *Visualization Analysis and Design*. A.K. Peters visualization series. Peters/CRC Press, 2014.
- [MVT16] Belgin Mutlu, Eduardo Veas, and Christoph Trattner. Vizrec: Recommending personalized visualizations. *ACM Transactions on Interactive Intelligent Systems*, 6(4):31:1–31:39, November 2016.
- [MVT17] Belgin Mutlu, Eduardo Veas, and Christoph Trattner. Tags, titles or q&as?: Choosing content descriptors for visual recommender systems. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT '17*, pages 265–274, New York, NY, USA, 2017. ACM.
- [MVTS15a] Belgin Mutlu, Eduardo Veas, Christoph Trattner, and Vedran Sabol. *Towards a Recommender Engine for Personalized Visualizations*, pages 169–182. Springer International Publishing, Cham, 2015.
- [MVTS15b] Belgin Mutlu, Eduardo Veas, Christoph Trattner, and Vedran Sabol. Vizrec: A two-stage recommender system for personalized visualizations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion, IUI Companion '15*, pages 49–52, New York, NY, USA, 2015. ACM.
- [Naz16] Kawa Nazemi. Adaptive Semantics Visualization. (November):422, 2016.
- [NKA11] F. Matsatsinis Nikolaos, Lakiotaki Kleanthi, and Tsoukiás Alexis. Multicriteria user modeling in recommender systems. *IEEE Intelligent Systems*, 26:64–76, 2011.

- [NRB<sup>+</sup>13] Kawa Nazemi, Reimond Retz, Jürgen Bernard, Jörn Kohlhammer, and Dieter Fellner. Adaptive semantic visualization for bibliographic entries. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8034 LNCS(PART 2):13–24, 2013.
- [OHBR15] Thomas Orgel, Martin Hffernig, Werner Bailer, and Silvia Russegger. A metadata model and mapping approach for facilitating access to heterogeneous cultural heritage assets. *International Journal on Digital Libraries*, 15(2-4):189–207, 2015.
- [PS13] Denis Parra and Shaghayegh Sahebi. Recommender systems: Sources of knowledge and evaluation metrics. In *Advanced Techniques in Web Intelligence-2*, volume 452 of *Studies in Computational Intelligence*, pages 149–175. Springer Berlin Heidelberg, 2013.
- [RB01] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB The Very Large Data Bases Journal*, 10(4):334–350, December 2001.
- [Rij74] C.J. Van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.
- [RKAES13] Majdi Rawashdeh, Heung-Nam Kim, JihadMohamad Aljaam, and Abdulmotaleb El Saddik. Folksonomy link prediction based on a tripartite graph for tag recommendation. *Journal of Intelligent Information Systems*, 40(2):307–325, 2013.
- [RKR06] Philippa Rhodes, Eileen Kraemer, and Bina Reed. VisIon: an interactive visualization ontology. *ACM-SE 44: Proceedings of the 44th annual Southeast regional conference*, pages 405–410, 2006.
- [RL95] Penny Rheingans and Chris Landreth. Perceptual principles for effective visualizations. In Georges Grinstein and Haim Levkowitz, editors, *Perceptual Issues in Visualization*, pages 59–73, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.
- [RN07] Francesco Ricci and Quang Nhat Nguyen. Acquiring and revising preferences in a critique-based mobile recommender system. *IEEE Intelligent Systems*, 22(3):22–29, May 2007.
- [RRSK10] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. *Recommender Systems Handbook*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

- [SAR08] Gao Shu, Nick J. Avis, and Omer F. Rana. Bringing semantics to visualization services. *Advances in Engineering Software*, 39(6):514–520, 2008.
- [SCC13] Ben Steichen, Giuseppe Carenini, and Cristina Conati. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. *Iui*, pages 317–328, 2013.
- [SDKP06] Ahmed Seffah, Mohammad Donyaee, Rex B. Kline, and Harkirat K. Padda. Usability measurement and metrics: A consolidated model. *Software Quality Control*, 14(2):159–178, June 2006.
- [SFHS07] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. The adaptive web. chapter Collaborative Filtering Recommender Systems, pages 291–324. Springer-Verlag, Berlin, Heidelberg, 2007.
- [SK09] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:4:2–4:2, jan 2009.
- [SKK10] Markus Strohmaier, Christian Koerner, and Roman Kern. Why do users tag? detecting users’ motivation for tagging in social tagging systems. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.
- [SPUP02] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’02, pages 253–260, New York, NY, USA, 2002. ACM.
- [SSK<sup>+</sup>14] Florian Stegmaier, Christin Seifert, Roman Kern, Patrick Höfler, Sebastian Bayerl, Michael Granitzer, Harald Kosch, Stefanie Lindstaedt, Belgin Mutlu, Vedran Sabol, Kai Schlegel, and Stefan Zwicklbauer. *Unleashing Semantics of Research Data*, pages 103–112. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.
- [Sta04] Gerry Stahl. What we know about cscl and implementing it in higher education. chapter Building Collaborative Knowing: Elements of a Social Theory of CSCL, pages 53–85. Kluwer Academic Publishers, Norwell, MA, USA, 2004.
- [STH02] C. Stolte, D. Tang, and P. Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–14, 2002.

- [STV<sup>+</sup>14] Vedran Sabol, Gerwald Tschinkel, Eduardo E. Veas, Patrick Höfler, Belgin Mutlu, and Michael Granitzer. Discovery and visual analysis of linked data for humans. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 309–324, 2014.
- [TdSMS15] Gerwald Tschinkel, Cecilia di Sciascio, Belgin Mutlu, and Vedran Sabol. The recommendation dashboard: A system to visualise and organise recommendations. In *19th International Conference on Information Visualisation, IV 2015, Barcelona, Spain, July 22-24, 2015*, pages 241–244, 2015.
- [TKL15] Christoph Trattner, Dominik Kowald, and Emanuel Lacic. Tagrec: Towards a toolkit for reproducible evaluation and development of tag-based recommender algorithms. *ACM Special Interest Group on Hypertext and the Web, SIGWEB Newsl.*, (Winter):3:1–3:10, February 2015.
- [TM04] Melanie Tory and Torsten Müller. Rethinking visualization: A high-level taxonomy. *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*, pages 151–158, 2004.
- [TVS<sup>+</sup>17] Thomas Thiele, André Calero Valdez, Sebastian Stiehm, Anja Richert, Martina Ziefle, and Sabina Jeschke. Augmenting research cooperation in production engineering with data analytics. *Production Engineering*, 11(2):213–220, Apr 2017.
- [vE02] Jörg von Engelhardt. *The Language of Graphics: A Framework for the Analysis of Syntax and Meaning in Maps, Charts and Diagrams*. Yuri Engelhardt, 2002.
- [VFM13] Martin Voigt, Martin Franke, and Klaus Meissner. Using expert and empirical knowledge for context-aware recommendation of visualization components. *IARIA International Journal on Advances in Life Sciences*, 5(1):27–41, 2013.
- [VGD<sup>+</sup>14] Katrien Verbert, Sten Govaerts, Erik Duval, Jose Luis Santos, Frans Van Assche, Gonzalo Parra, and Joris Klerkx. Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing*, 18(6):1499–1514, Aug 2014.
- [VHS<sup>+</sup>15] Manasi Vartak, Silu Huang, Tarique Siddiqui, Samuel Madden, and Aditya Parameswaran. Towards Visualization Recommendation Systems. *1st Workshop on Data Systems for Interactive Analysis (DSIA)*, (Section 2):1, 2015.

- [VMdS<sup>+</sup>15] Eduardo E. Veas, Belgin Mutlu, Cecilia di Sciascio, Gerwald Tschinkel, and Vedran Sabol. Visual recommendations for scientific and cultural content. In *IVAPP 2015 - Proceedings of the 6th International Conference on Information Visualization Theory and Applications, Berlin, Germany, 11-14 March, 2015.*, pages 256–261, 2015.
- [VMPP14] Manasi Vartak, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. Seedb: Automatically generating query visualizations. *Proc. VLDB Endow.*, 7(13):1581–1584, August 2014.
- [VWV<sup>+</sup>07] Fernanda B. Viegas, Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.
- [War12] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3 edition, 2012.
- [WMA<sup>+</sup>16] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, 2016.
- [Wol15] Michael Wollowski. Search and inference with diagrams. (March), 2015.
- [WSP<sup>+</sup>06] William Wright, David Schroh, Pascale Proulx, Alex Skaburskis, and Brian Cort. The sandbox for analysis: Concepts and methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06*, pages 801–810, New York, NY, USA, 2006. ACM.
- [Yao03] Y. Y. Yao. *Information-Theoretic Measures for Knowledge Discovery and Data Mining*, pages 115–136. Springer Berlin Heidelberg, 2003.
- [ZLZK07] Xianjun Sam Zheng, James j W. Lin, Salome Zapf, and Claus Knapheide. Visualizing user experience through "perceptual maps": Concurrent assessment of perceived usability and subjective appearance in car infotainment systems. In *Proceedings of the 1st International Conference on Digital Human Modeling, ICDHM'07*, pages 536–545, Berlin, Heidelberg, 2007. Springer-Verlag.