UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

FABIAN ERNESTO COLQUE ZEGARRA

VUGA: A Visual User Group Analytics System for User Data

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Prof. Dr. João Luiz Dihl Comba

Porto Alegre
February 2019

CIP — CATALOGING-IN-PUBLICATION

Colque Zegarra, Fabian Ernesto

VUGA: A Visual User Group Analytics System for User Data / Fabian Ernesto Colque Zegarra. – Porto Alegre: PPGC da UFRGS, 2019.

64 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2019. Advisor: João Luiz Dihl Comba.

I. Dihl Comba, João Luiz. II. Título.

“An inventor is simply a fellow who doesn’t take his education too seriously.”

— CHARLES F. KETTERING

ACKNOWLEDGEMENTS

I deeply appreciate my advisor João Comba for all the teachings during my masters. I thank all the people who helped me during my stay in Brazil, especially my friends Jose Luis, Jorge, Abel, Lizeth and Juan. Last but not least, I thank my family, my parents Benancio and Dina, my uncles Lourdes and Albino for supporting me since my graduation until now.

ABSTRACT

Along with the constant growth in the size of user data, the need to understand the behavior of user groups increases. The study of user data is known as User Group Analytics(UGA). UGA is used to give support in making the best decisions quickly and with greater credibility. In addition to addressing certain peculiarities such as noise and the dispersion of data, UGA is helping the scientific community to carry out studies and experiments on large-scale population. However, it also helps ordinary users to make more routine decisions either with people who think in a similar way as well as in a different way. The ability to explore and compare information of interest is within the core of UGA. While there are already automated systems that can identify and suggest potentially interesting groups, there is a need to filter results and improve the parameters as a user explores the data needs to finally provide a visual interface. Consequently, the combination of a visual and analytical interface with the opportunity for reevaluation filters and feedbacks are known as Visual Analytics (VA). In the present work, we describe VUGA, a VA system that explores groups of user data and improves other approaches based only on automated procedures.

Keywords: Visual Analytics. Visualization Information. Exploration and Discovering of Groups.

VUGA: Um sistema de análise visual de grupo de usuários para dados de usuários

RESUMO

Juntamente com o constante crescimento no tamanho dos dados do usuário, aumenta a necessidade de entender o comportamento dos grupos de usuários. Esse fato é conhecido como User Group Analytics(UGA), a UGA é usada para dar suporte na tomada das melhores decisões rapidamente e com maior credibilidade. Além de abordar certas peculiaridades como ruído e dispersão de dados. A UGA está ajudando a comunidade científica a realizar estudos e experimentos em população em larga escala. No entanto, também ajuda os usuários comuns a tomar decisões mais rotineiras com pessoas que pensam de maneira semelhante e de maneira diferente. A capacidade de explorar e comparar informações de interesse está dentro do núcleo da UGA. Embora já existam sistemas automatizados que possam identificar e sugerir grupos potencialmente interessantes, mas também é necessário filtrar os resultados e melhorar os parâmetros à medida que o usuário explora as necessidades nos dados para finalmente fornecer uma interface visual. Conseqüentemente, a combinação de uma interface visual e analítica com uma oportunidade para filtros de reavaliação e feedbacks é conhecida como Visual Analytics (VA). No presente trabalho, descrevemos o VUGA, um sistema VA que explora grupos de dados de usuários para melhorar abordagens baseadas apenas em procedimentos automatizados.

Palavras-chave: Visual Analytics. Visualization Information. Exploration and Discovering of Groups..

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

VA	Visual Analytics
t-SNE	t-Distributed Stochastic Neighbor Embedding
PCA	Principal Component Analysis
MDP	Multi-Dimensional Projection
MDS	Multi-Dimensional Scaling
Lamp	Local Affine Multidimensional Projection

CONTENTS

1 INTRODUCTION

The increase of information on user data has grown significantly in the last decade. This fact has led us to the need to understand the behavior of groups. Scientists and non-scientists can understand the influence of these groups at the moment of discovering patterns or identifying collective behaviors. Some examples are user data consisting of product reviews, medical records, scientific publications, and retail store receipts. These data can be described from a general view as a combination of demographics (e.g., age, gender, occupation) and actions (e.g., movies, books or even medical treatments). The identification of group behavior in these datasets is based on the ability to explore the user record space and add their demographic characteristics and behaviors. Currently, there are automated systems to identify and suggest potential and interesting group behaviors, usually based on artificial intelligence techniques that receive standard input and return fixed information (Wang; Cao; Chi, 2015), (Rein et al., 2017). While a more personalized search will require a set of filters applied to the data through a graphical interface, the output will be a set of refined data based on human analysis decisions. The combination of a visual interface with analysis, with the ability to filter information and the ability to have feedback is referred to as Visual analytics (VA) (Cheng et al., 2016). In this work, we present a VA approach called VUGA (Visual User Group Analytics), the first VA system to explore groups of user data.

This VUGA system is composed of three main components: Discovery, Exploration, and Visualization. These three components help to understand the execution and use of the system. Discovery is the process through which users observe, build and find groups. The output in this component is a set of groups with similar characteristics between the new groups and the original group. Inside this component we can see the development of algorithms that seek to find the best answers to user data groups. In the Exploration component, the interaction between the users and the system data is allowed in order to create knowledge and acquire the most significant possible learning during the duration of this user interaction with the system. The third component, Visualization, allows a detailed inspection of the information available to the user. Such components give support to a better elaborated and adequate response for a later purpose based on exploring and discovering pipelines. It is necessary to build a all-in-one system using the three components.

Additionally, a user group analysis system must provide an adequate environment

for different levels of user experience. In the case of a novice user, we will have those people who are interested in completing daily tasks such as finding a movie or a book. For that, they need first to find other people like them to exchange between a user-level view and a group level view. It is exploring individual and collective preferences to reach a decision. On the other hand, in the domain of experts tend to observe to validate some assumption in the data they are working (which is called "Confirmation Analysis"). For example, they want to verify if most people prefer Drama films or Comedy. For that, they first need to obtain a holistic view of statistics and data distributions associated with data. Finally, data scientists are interested in testing their algorithms and seek to facilitate the implementation and connection of different components in order to have a starting point for data analysis. All these types of users have the same objective that is to explore the data holistically. The objective of this work is to offer an integrated environment with which all types of users can comfortably complete their tasks.

The need for systems to quickly build groups and filter them at that moment along all their dimensions generates the challenge that it is useful for different types of users. For a scientist, these characteristics can be valuable when analyzing new concepts and standards. For an expert, it will be useful when extracting relevant information in a more simplified way but with the same value. While for a more inexperienced user, it will be helpful to find information that they already knew but that is complicated to obtain with other methods.

The need to obtain different statistics associated with a group of interest has some advantages. For a scientist, these statistics are valuable to show hypothesis and conclusions. For an expert using this feature will be helpful in order to have a broader view of the field and consequently being able to make decisions. While for a novice user, this feature will be helpful in order to have complementary information.

The need for discovering groups similar to other groups of interest can be useful in a different way. Scientists are interested in being able to discover new patterns within the data. For expert users, discovering groups will be more useful when drawing market and business conclusions. While for novice users, this information will be useful for finding information related to everyday activities.

1.1 Contributions

This work is a visual analytics system that integrates several components: Projection of multidimensional data, visualization, and filtering of multidimensional data, the discovery of groups and similarity among them, visualization of similarity between groups. All these components come together to create a VA system that explores, visualizes and discovers new information in different levels of user experience. In the next list are described the contributions of this work:

Building Groups The ability to build and discover groups and filter on-the-fly along multiple dimensions;

Statistics Views A coordinated view of various statistics associated with a group of interest based on similarity;

Similarity Using the similarity, we can create a group with specific characteristics and discover new similar groups with associated characteristics;

Several Use Cases We introduce a set of scenarios that illustrate the power of VUGA.

1.2 Document organization

This work is organized as follows. In Chapter 2 we describe related work to VUGA, visualization, and multidimensional data. Chapter 3 describes the multidimensional data and how this data was modeled for a VA system. In Chapter 4 different visual components of the VUGA are introduced and explained. In Chapter 5, we explain the process of discovering new groups. Evaluation with use cases and discussion with user evaluation about VUGA is developed in Chapter 6. Finally, we summarize our conclusions and present new ideas for future work.

2 RELATED WORK

In this chapter, we describe related work divided into sections according to each topic. Initially, we give a historical analysis of the exploration and analysis of data. After, we divide previous work into two parts: visualization of high dimensional data and exploration and analysis of user groups. All these related works were used to choose solutions in each part of our work.

2.1 Visualization and Exploration of High Dimensional Data

In recent works about user data, the visualization, exploration, and analysis of data have become a fundamental topic of study for understanding the different challenges about user data. A tutorial about the different techniques in the exploration of data is given in by Idreos and colleagues (IDREOS; PAPAEMMANOUIL; CHAUDHURI, 2015). They discuss emerging developments in the database area and how data exploration allows for creating new interfaces. In table 2.1, we summarize their taxonomy of data management for different data exploration tasks. For example, the relation of data visualization and exploration interfaces in the creation of user interfaces.

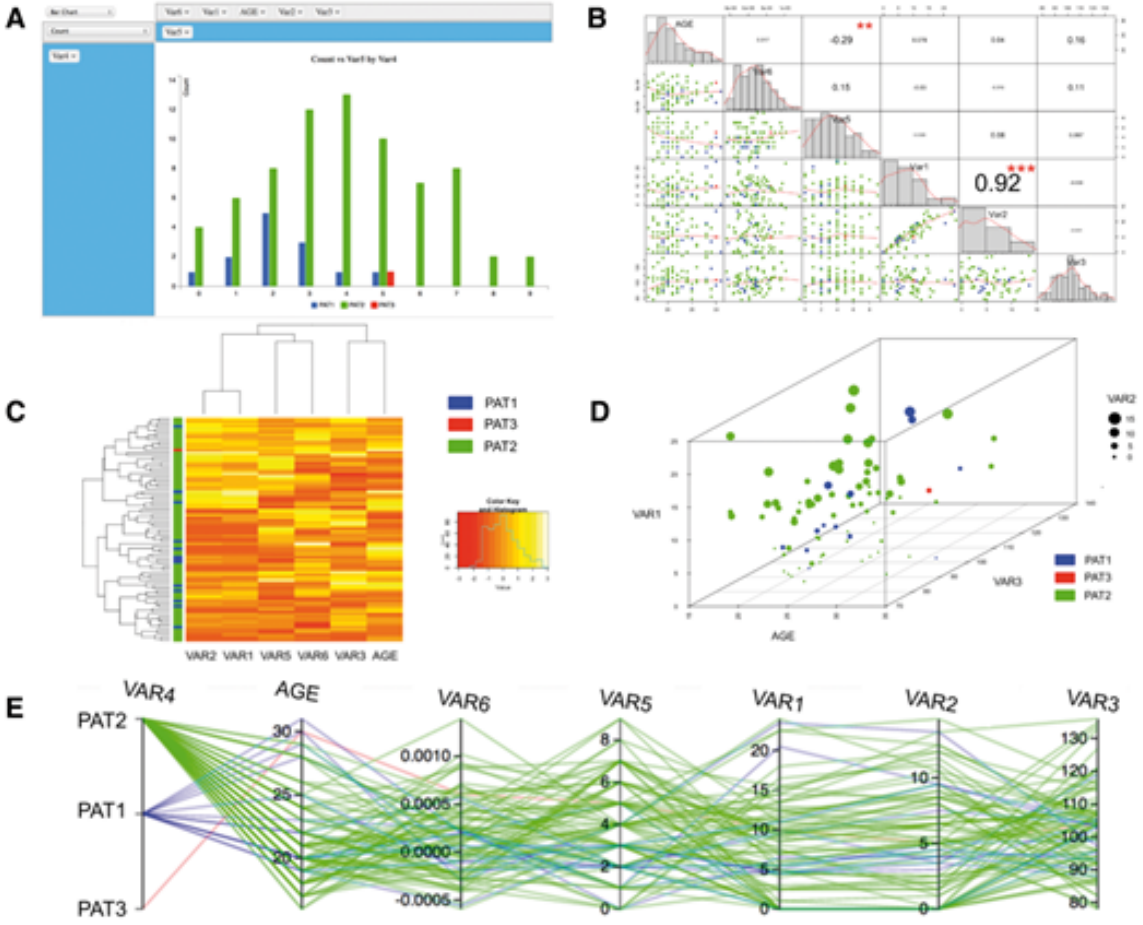
Several works describe the exploration of multidimensional data, and the need to understand what information can be extracted from this type of data (JO et al., 2017; DIMITRIADOU; PAPAEMMANOUIL; DIAO, 2016). Before the visualization or exploration of massive data, data needs to be pre-processed and prepared for analysis and to answer queries. One approach of pre-processing is used in SwiftTuna (JO et al., 2017), a holistic

Table 2.1: Data management research for data exploration.

User Interaction	**Data Visualization**	Visual Optimizations	Visualization Tools	
	Exploration Interfaces	Automatic Exploration	Assisted Query Formulation	Novel Query Interfaces
Middleware	**Interactive Perfomance Optimizations**	Data Prefetching	Query Approximation	
Database Layer	**Indexes**	Adaptive Indexing	Time Series	Flexible Engines
	Data Storage	Adaptive Loading	Adaptive Storage	Sampling

Fonte: Idreos, Papaemmanouil and Chaudhuri (2015, p. 279)

Figure 2.1: Examples of common visualization techniques for multidimensional data.

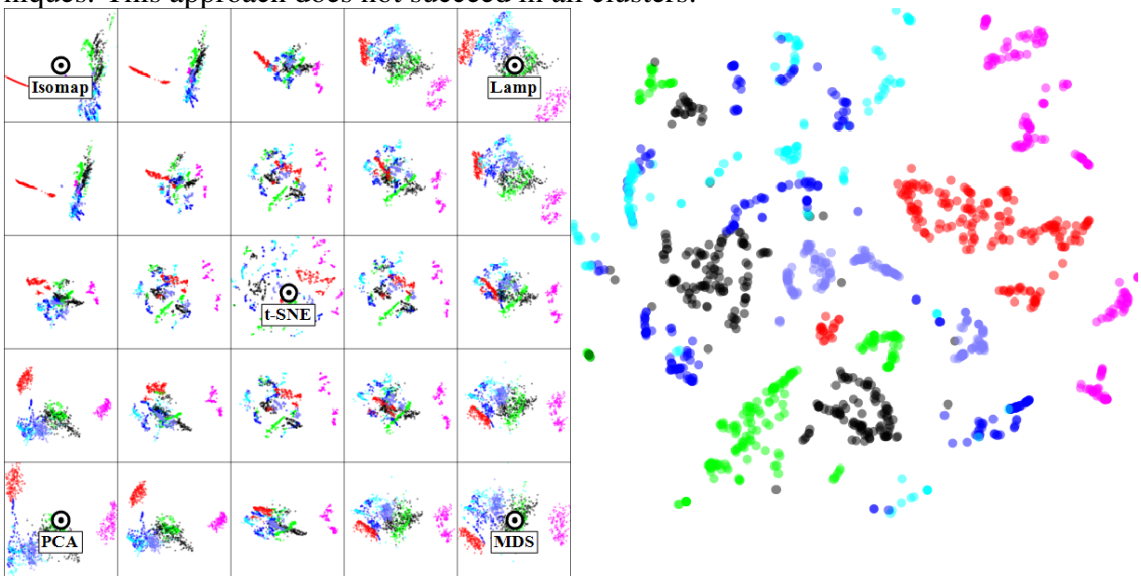


Fonte: Dunn Jr et al. (2017, p. 3)

system that streamlines the process of searching for visual information in multidimensional data. In that work, the authors developed a novel visualization technique (tailed charts) to facilitate large-scale multidimensional data exploration. The tailed charts were divided into two types: the tailed dot plots and tailed gradient plots. These two variants help to improve the understanding of the frequency histograms.

(DUNN JR et al., 2017) reviewed the current state of the existing tools for the visualization of multidimensional data in transactional research platforms. An example of their proposed visualization is shown in Fig. 2.1. In Fig.2.1.**a** is shown a dynamic pivot table, which allows a variety of visualization for multidimensional data, like changing between several bar charts for each dimension. In Fig.2.1.**b**, a correlation matrix is used when there is dependences between multiple variables. For this type of visualization, we can use different methods for correlation analysis, like the Pearson parametric correlation test. The heatmap in Fig.2.1.**c** shows a visualization technique helpful to extract a global view of the numerical data. However, it also needs more pre-processing, such as data

Figure 2.2: Segmentation in several clusters using a table with different projection techniques. This approach does not succeed in all clusters.



Fonte: Kruiger et al. (2017, p. 13)

normalization and permutation of rows and columns to place similar values and discover clusters. The 3D ScatterPlot is shown in Fig.2.1.**d**, with the data in a 3D plane. Finally, in Fig.2.1.**e** we have parallel coordinates, which is one of the most known and is used to display a larger number of dimensions. This technique allows for discovering data clusters.

In (KRUIGER et al., 2017), the focus is on the combination of visualization techniques and methods for extracting information in multidimensional data. One of their approaches was to develop a tool which has a table with five projection techniques(Isomap, Lamp, t-SNE, PCA, MDS) as shown in Fig. 2.2. The goal of this table is to allow changes in the projection of the data on the right side of the figure. This tool also allows to set a subset of the data and prevent them from being modified with the interaction of the table. The final visualization is a set of different clusters, as shown in the figure. The work of Kruiger et al.2.2 was useful to decide what type of projection we will use in the development of our work.

The need for more complex visualizations leads to the development of more advanced techniques. Bahador Saket (SAKET et al., 2017) proposes a paradigm for visualization by demonstration, which allows showing the incremental changes in the data and their visualization representation. This paradigm involves the interaction and combination between the user and the system to reach potential recommendations. The way to show the information was very useful to choose the type of visual chart to use. In Fig. 2.3

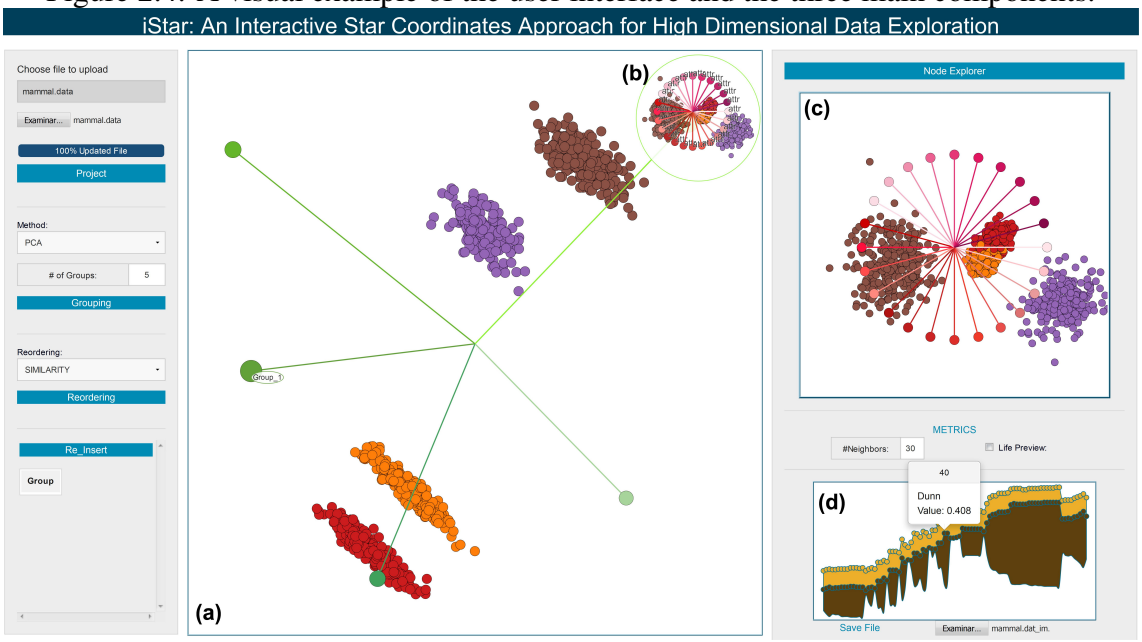
Figure 2.3: A VisExemplar user interface with the three main components.

Fonte: Saket et al. (2017, p. 4)

we show three parts of the system. The first one is a *ThinkBoard*, which allows users to construct their demonstrations through direct manipulation in the visualization. Using the interaction in the ThinkBoard, it is possible to click in a data point and see in the *Details View* when more information about the data. Also, in the *Recommendation Gallery* the visual representation transformation is displayed to summarize the interaction between the different components of the system.

Another approach for the visualization of multidimensional data and analysis of groups was developed in iStar(ZANABRIA; NONATO; GOMEZ-NIETO, 2016). Their approach uses star coordinates to discover new clusters. Such an approach allows grouping attributes which are useful to understand information, as shown in Fig. 2.4. The system allows operations such as scaling, rotation, union, separation, removal, re-insertion and position tuning. Such operations allow to explore and analyze information in a representation of clusters into the star coordinates, thus allowing to play with each border and separate the data in different ways. The figure also displays the different components used for the interaction. On the left side, we can see a menu, where it is possible to configure the scenario, for example, the PCA algorithm for dimensionality reduction. The example in the figure shows five groups differentiated by color and position in a 2D space (Fig. 2.4.**a**). In the projection they defined a node preview(Fig. 2.4.**b**) to allow select data into start coordinates. This selection can be visualized in a node explorer(Fig. 2.4.**c**), where it can be scaled, rotated and reflected in the central visualization. Finally, there is a visu-

Figure 2.4: A visual example of the user interface and the three main components.



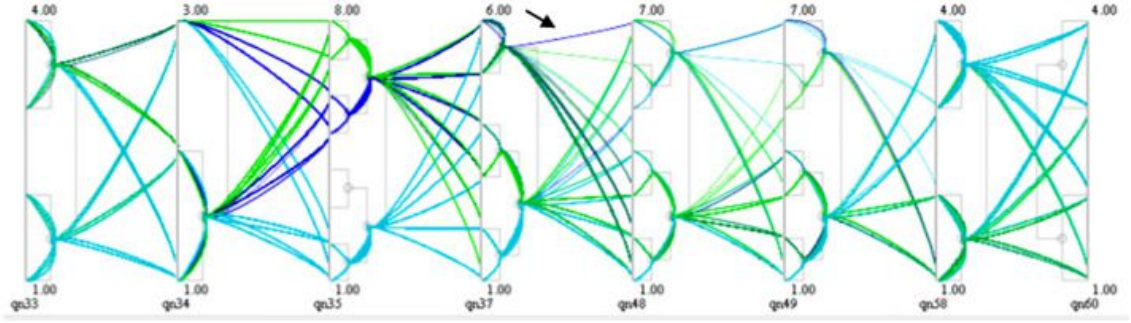
Fonte: Zanabria, Nonato and Gomez-Nieto (2016, p. 5)

alizer for quality(Fig. 2.4.**d**), where it is displayed the information about the evolution of the quality by each previous interaction using a stacked graph metaphor. This work gave us some ideas on how to explore the data in multidimensional spaces and the way to use the interaction human-computer to explore the data.

Another challenge in the visualization of high dimensional datasets is s large number of dimensions. For example Huang and Zhang in their work about the future in Computer Science(HUANG; HUANG; ZHANG, 2016; ITOH et al., 2017) used parallel coordinates to explore and interact with large datasets. Both works used parallel coordinates as a graph, with interaction for selection data. In the first work, Huang and colleagues presented a new technique for improving the selection interaction using parallel coordinates. Figure Fig. 2.5(**a**) shows the parallel coordinates modified to use a dendrogram for each dimension. This dendrogram is the result of a matrix design by the dimensions, where the concurrence of the data determine how the dendrogram will be obtained. As a result, the visualization becomes easier to interpret.

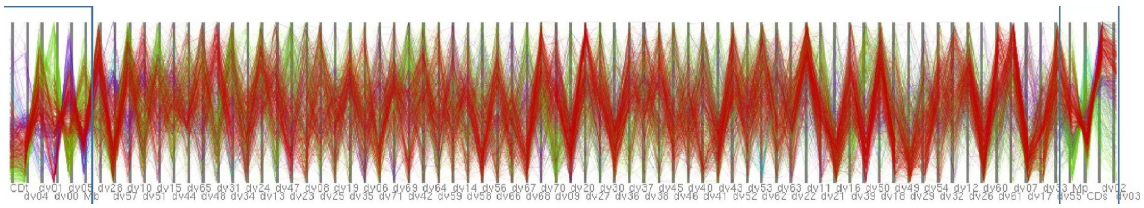
Itoh et al. (ITOH et al., 2017) use parallel coordinates for exploring high dimensional data in a low-dimensional space. The idea is to build a graph-structure with all correlations of the dimension, and the selection of dimensions to explore data in a low-dimension space. Figure 2.5(**b**) shows parallel coordinates of a large dataset. In principle, the general view of all dimensions is useful to find certain patterns, but it is difficult to explore and get insights from the data. To overcome this problem, they allow selecting

Figure 2.5: Parallel Coordinates used for visualization of high dimensional data.



(a) Parallel Coordinates with dendrograms

Fonte: Huang, Huang and Zhang (2016, p. 5)



(b) Parallel Coordinates with filters

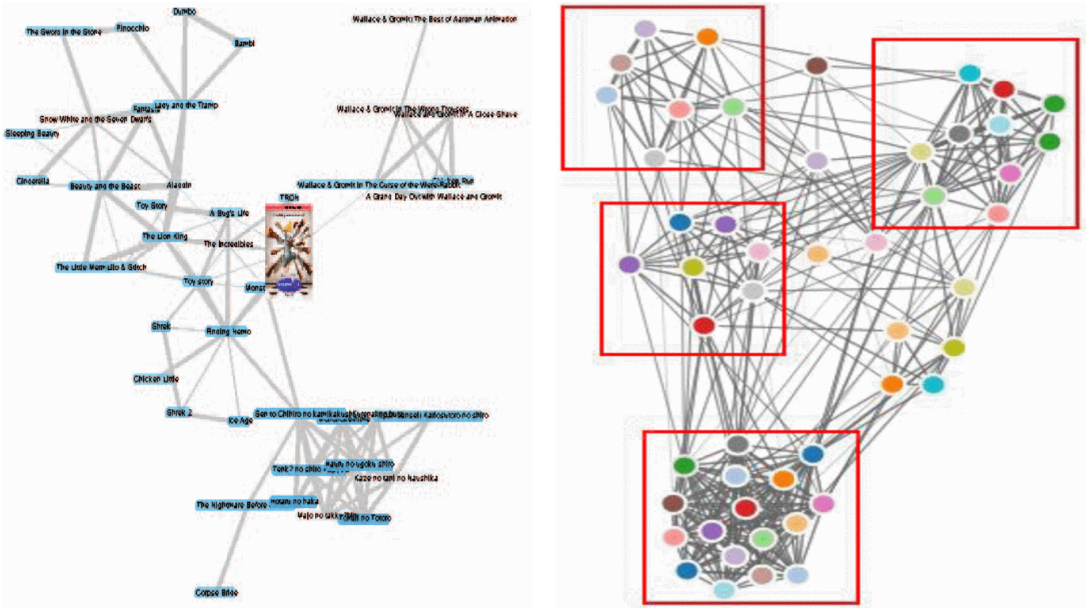
Fonte: Itoh et al. (2017, p. 5)

certain regions and zooming in to inspect details of the data. Colors are also used to differentiate between certain attributes. Additionally, after selecting a region, it is possible to see the graph with all the data and its relationships based on the dimensions and similarities.

Many use cases result from applying the different techniques in high data visualization dimensions. One of the most used and vastly explored enough in recent years are recommendation systems. Das et al. (DAS et al., 2015) proposed a way to handle scalability for large data using a classification based on recommendation algorithms called Weight Voronoi decomposition. The idea was to divide the space of users into smaller regions using the location as a measurement, and applying the algorithm for each region. The datasets used in their work are also used in this work. The Weighted Voronoi decomposition was useful to reduce the time of response in the recommendation system.

As a final comment, we mention the area of interaction between the data visualization and the analyst. The use of most current visualization techniques includes interaction with the user. This fact means that the interaction factor helps and improves the process of obtaining results. Dimitriadou et al. (DIMITRIADOU; PAPAEMMANOUIL; DIAO, 2014) present a framework for the interactive data exploration(IDE). This work introduces a variant called AIDE(Automatic Interactive Data Exploration) framework to guide users to interactively explore data, and feedback as a basis to improve collecting new informa-

Figure 2.6: Grouping of data based in correlation of attributes. **(a)** Fonte: Huang, Huang and Zhang (2016, p. 5), **(b)** Fonte: Huang, Huang and Zhang (2016, p. 5)



(a) Grouping movies according to their attributes (b) Improved groups using force-oriented layout

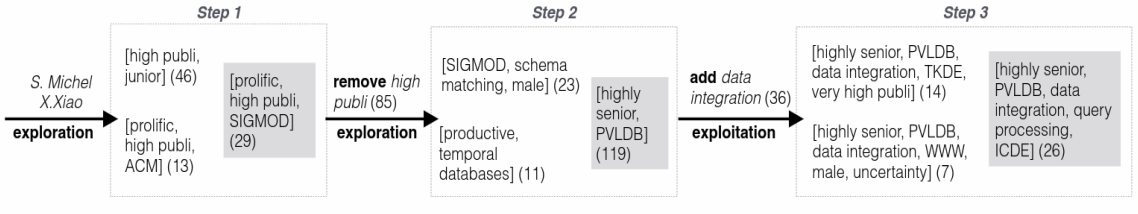
tion. As a result, the interaction inside the visualization is a part important for discovering new patterns, new information and the constant improvement of the quality of the data.

2.2 Exploration and Analysis of User Groups

The advances in the visualization and exploration of high-dimensional data brought many exciting applications. In this section, we describe the exploration and analysis for user groups as one of the recent problems in knowledge engineering and visual analysis.

There are different proposals for creating data groups. Some approaches use visualization and data mining algorithms for the discovery of new information. QIAN et al. (QIAN et al., 2016) propose a system to support recommendation systems for a movie dataset. The interface uses a multivariate data visualization combined with several views. They display a graph, where each node represents a data(user), and edges describe the correlation of attributes among users. They use a view to display the details of a subregion selected from the main graph and an image to display the attributes of the data. Fig. 2.6.**(a)** shows the connection between movies based on their attributes to identify groups of data. This idea was later improved by Dong with his work about interactive design on recommender system(DONG; CHENG; MIN, 2017), where a force-based algorithm was used to indicate how close or how far two elements are. Forces are directly proportional

Figure 2.7: Discovering Users in a Program Committee with IUGA



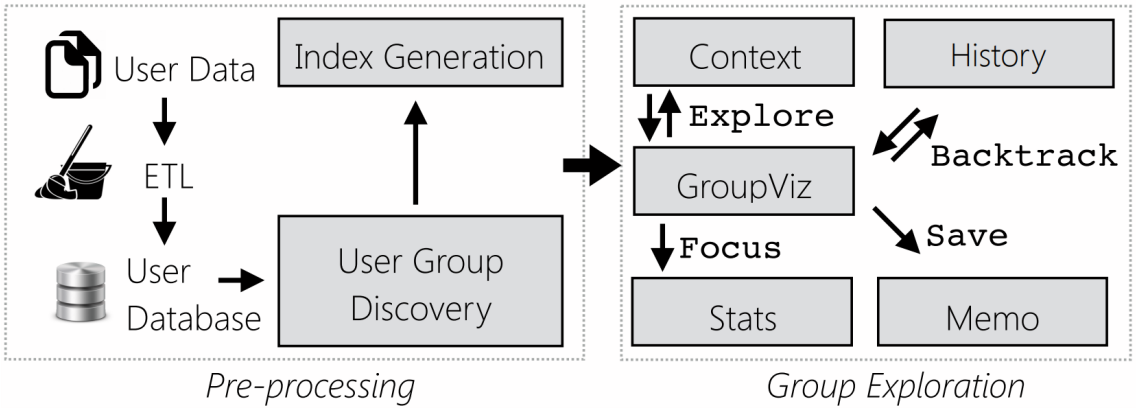
Fonte: Omidvar-Tehrani, Amer-Yahia and Termier (2015, p. 2)

to the total similarity between two points (element). They improved the attribute-oriented algorithm to visualize the media project data and also presented a study for collaborative filtering. Fig. 2.6 **(b)** shows an example of how grouping elements using a force of similarity, additionally is used color as an attribute encoded.

A different approach for the discovery of new groups of users is described in (HU; YAO; CUI, 2014), where a probabilistic unified model called GrosToT (Group Specific Topics over time) is proposed. This model allows understanding user groups together with temporary topics at the same time. Also, the model verifies the relationship (event) that exists between the user group and the temporal dynamics, and how it varies or remains stable as the data arrive.

The analysis of groups becomes important in other contexts. For example, with the increase in information through different social media on the web, the interest in analyzing that data increases. As we have already commented before, these studies are quite well studied in population studies, as well as in recommendation systems, among others. For example, Mazumdar(MAZUMDAR; PETRELLI; CIRAVEGNA, 2014; FEKETE; PLAISANT, 2002) describes the Semantic web, and how the exploration of linked data is an open problem. They proposed to consider the points of view of each user and developed visualization to explore the relations between data. For data analysis Behrooz(OMIDVAR-TEHRANI; AMER-YAHIA; TERMIER, 2015) proposes a framework (IUGA), which allows producing representative groups for a set of users that have properties in common. This framework uses a set of primitives that allow the discovery of new groups of users; those primitives are for analysis (add and remove users), as well as for exploration (explore, exploit). The term *explore* refers to finding new users out of the set, while the term *exploit* refers to finding new users into the set of users. Figure 2.7 shows an example of user discovery in the problem of the formation of a program committee. In their proposal a group is first initialized with a user, properties are iteratively removed from the dataset to explore new data, and finally, a new property is added before calling the exploit operation

Figure 2.8: Architecture of VEXUS



Fonte: Amer-Yahia et al. (2017, p. 2)

to find new users within the existing set.

Additionally, in the work VEXUS (AMER-YAHIA et al., 2017), it was presented an idea to incorporate a visual functionality for the exploration and analysis of data using IUGA as a generator of new groups of users. In this proposal, the interaction of the analyst was improved, with the support of the visualization of the explore and exploit operations. We added filters to allow the selection of users. The interface allows to store the history of the interaction, and therefore to be able to return to an exact point of the interaction at any time. Additionally, this proposal allows interacting with large amounts of data and different ways of interaction with the data to find the best groups of users. This approach follows the architecture showed in Fig. 2.8, where it is defined two main parts: pre-processing, which is responsible for managing the storage of data through indexing and processing of the IUGA algorithm, and exploration of groups.

So, in summary, we reviewed some work with some contributions in how visualize multidimensional data. Other works showed us how interact with multidimensional data. Some works also showed us what type of visualization are ideal to show multidimensional data. Each one of that works gave us a lot of information and ideas in how develop our contributions to explore and visualize data in high dimension. In the second part of this revision, we can see the IUGA, a algorithm to discover groups since a set of static user groups. After this, we saw the VEXUS, a visual tool to use the IUGA, where a set of visualization are presented to interact and explore the set of groups. That works gave us our start point to develope our own ideas in how explore, visualize and discover user groups in real time.

3 DESGIN AND PRE-PROCESSING OF DATA

The architecture of VUGA (See Fig. 3.1) is composed of a preprocessing over the data (explained in this chapter) and a group exploration module (explained in the next two chapters). The preprocessing is a component that receives data in a given format. This component transforms and models the data to fit the needs of the VA system. In addition to the general demographics attributes, we add to the format the projection of an n−dimensional feature space that encodes desired user data used to express similarity.

We explain how we handled the high-dimensional data. We start with how we modeled the data in several dimensions and discussed the solution used to visualize multidimensional data in a 2D space. We intend to support as many types of data as possible from different areas of study. However, in this work, we focus on datasets containing reviews of users of some product (e.g., movies or books) and healthcare data.

3.1 Data Modeling

We use three datasets in this work: Movielens, BookCrossing, and Health dataset. These datasets were shared by our collaborators in this research (OMIDVAR-TEHRANI; AMER-YAHIA; TERMIER, 2015). These data sets went through a cleaning process without affecting their nature. That is, it eliminates some data that did not have relevance

Figure 3.1: VUGA architecture: It is formed by a pre-processing and a visual analytics interface to support interaction with user data and group exploration.

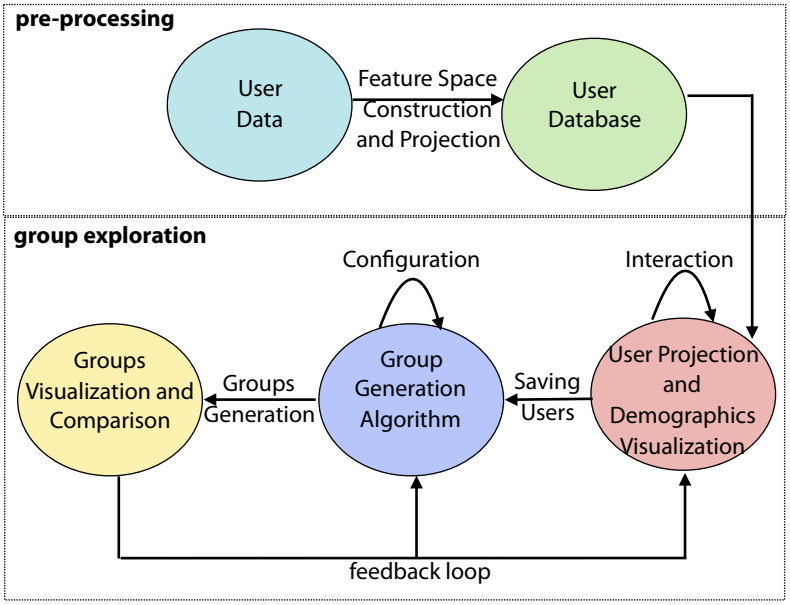


Table 3.1: Description of the Movielens dataset.

Movielens				
Item	**Description**			
User	User ID	Gender	Age	Occupation
Movie	Movie ID	Title	Genres	
Ratings	User ID	Movie ID	Rating	

Table 3.2: Description of the BookCrossing Dataset

Book Crossing					
Item	**Description**				
User	User ID	Gender	Age	Occupation	
Book	Book ID	Title	Author	Publisher	Year Publication
Ratings	User ID	Book ID	Rating		

Table 3.3: Description of the Health dataset information.

Health						
Item	**Description**					
User	User ID	Gender	Age	City	is_censored	is_dead
Action	Action ID	Title	Category			
Event	User ID	Action ID	Timestamp			

for the data, such as users with negative ages, users without a specific gender, among others.

The Movielens dataset has different versions on the web. The dataset contains user ratings to the movies they watched. This data has 1,000,209 ratings of 3,952 movies by 6,040 users. Users rate if a film is good or not through a metric (e.g., 1 star to 5 stars, the most common parameter used in favorite sites). Table 3.1 shows that we have the descriptions of the users, movies, and ratings that the users give to the movies. We used the genres of the movies as dimensions; there are 18 dimensions: Drama, Comedy, Action, Thriller, Sci-Fi, Romance, Adventure, Crime, War, Horror, Children, Animation, Mystery, Musical, Fantasy, Film-noir, Western, Documentary.

The BookCrossing dataset is similar to Movielens. It has 101,376 ratings of 46,380 books by 8,167 users. Since this dataset is bigger than the Movielens, it is more challenging. Table 3.2 shows the description of this dataset. There are users, books, and ratings for books by users. This figure was the original structure of the data; it was necessary to perform an additional search to extract the genre of books. There are 18 dimensions.

The format of this dataset is different from the previous datasets (Table 3.3). This dataset contains healthcare events that describe actions (medical treatments in a specific area) for different users. Currently, we have 1,000,000 medical consulting in 40 medical areas by 218,812 patients. In this case, we used the timestamp as the relation between users and actions. As in the previous dataset, we had to extract some new characteristics.

Table 3.4: Description of input dataset pattern

Pattern dataset			
Item	**Description**		
User	User ID	Gender	<Attributes List>
Object	Object ID	Description	<Attributes List>
Association	User ID	Object ID	Action
Dimensions	User ID	Dimensions List	

Therefore, we extracted the categories of medical areas in different treatments, resulting in 40 dimensions.

Since data can be in different formats, with different pieces of information, a previous pre-processing step is necessary before visualizing the data. As Kampars(KAMPARS; GRABIS, 2017) said: to get a better understanding of the data, we need to apply a step of data integration to improve the data-driven application. For our purpose, it was necessary to establish a single particular format.

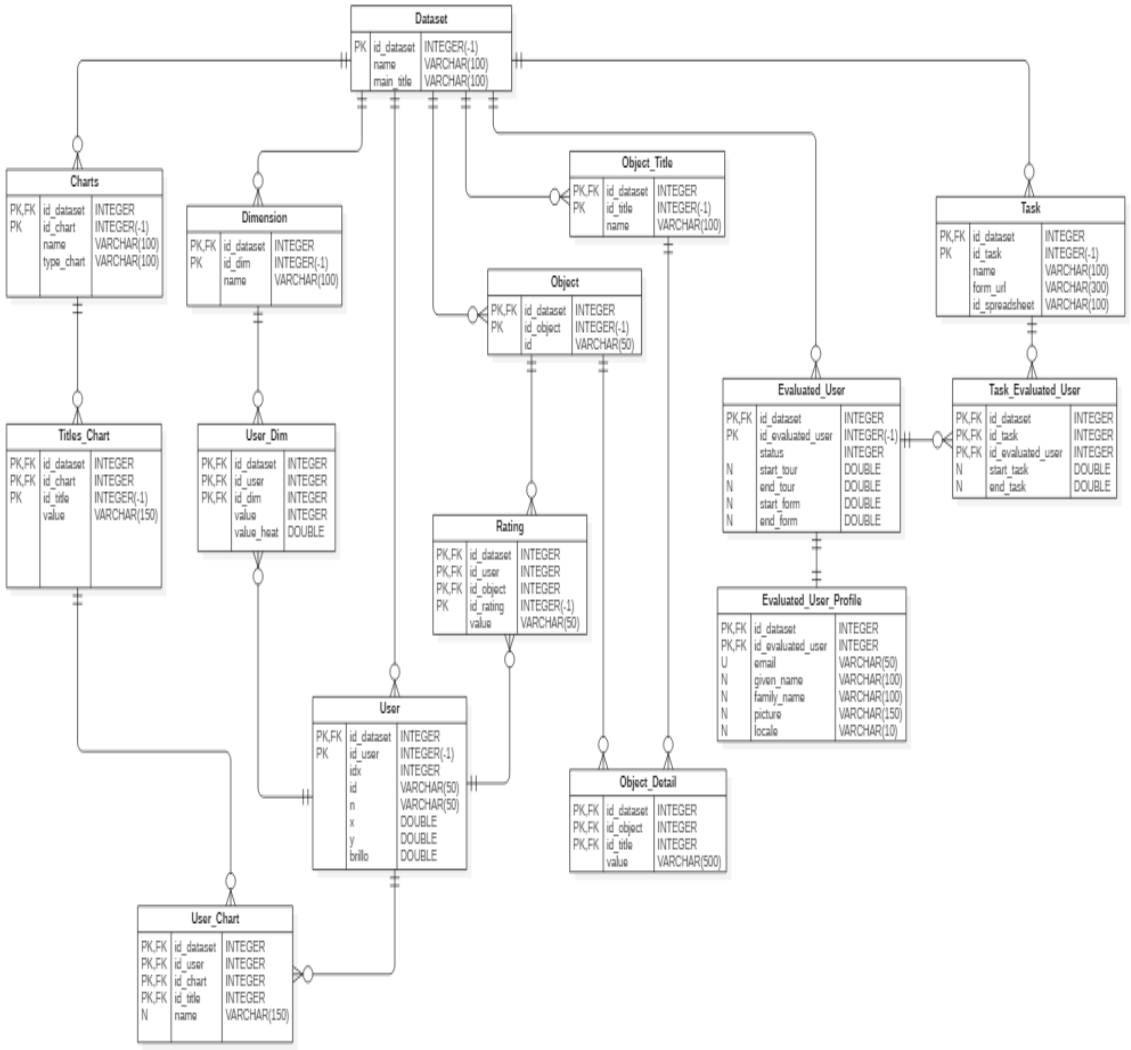
We defined four input files, shown in table 3.4. The first is the file reserved for the user; this file contains the demographic information of the user as well as some extra information according to the data set used. For example, in Movielens we have the information shown in Fig. 3.1 (User ID, Age, Occupation).

The second file is reserved for the object qualified by the user, which we call *object*. The dimensions are extracted from this file since we believe it is convenient to establish a relationship between the user and the Object. For example, in the Movielens dataset, we have the Movies file (Fig. 3.2), and its data are Movie Id, Title, Genre.

The third file is the qualification file which we call the *association file*. This file contains the information related to the union between the user file and the Object file. In the case of Movielens and BookCrossing we have that the ratio is the rating that a user gives to a Movie / Book, while in the dataset of Health, we have the relation as an event (medical consult) between a patient and an action (specific health area). In other words, this file has three pieces of information: the user ID, the Object ID and a value that represents the union between the user and the Object.

Finally, the fourth file contains information about the dimensions for each user, which consists of a matrix of users versus dimensions. As we mentioned before, the dimensions that we considered were extracted from the Object file and concatenated with each user. For example, in Movielens, we generate the dimension file by the number of movies of a particular genre that a user watches. In BookCrossing, we generate the file of dimensions based on the number of books of a specific genre that each user reads. Finally,

Figure 3.2: Database Model of VUGA.



in the HealthCare data, we generated the file based on the number of times a patient had a particular medical treatment. A summary of this pattern dataset can be observed in the table 3.4.

The way dimensions were defined for each dataset is explained in more detail in the next section. In order to store and use these datasets and other possible future datasets, we decided to use a database (See model in Fig. 3.2). This database has support for the structure previously explained. With the database we avoided enough redundancy in the data that was being stored. Response times accelerated considerably compared to when we used flat files. This DB is relation, in SQL using PostgreSQL (POSTGRESQL, 2018) as administrator. Also, it was necessary a server based on Tornado (TORNADO, 2016), a framework based on Python that allowed us to consume the information of the DB and display it in VUGA.

3.2 Data Embedding in nD

In this section, we detail the creation of multidimensional data and how was the normalization process of each dimension. Our model is based on the idea that each user has n dimensions that represent their characteristics. In order to define what is a dimension and what is the value that contributes to the dataset, it is important to review all the available data and observe the behavior of each one. The choice of dimensions in our data was dependent on the type of data (e.g., Numeric, Categorical, Ordinal) of each element.

We first discuss if a demographic attribute (e.g., Age, Gender, Occupation) of the Users file could be considered as dimensions for a user. One limitation is the fact that this information is categorical, which requires additional processing that increases the number of dimensions. For example, Bertjan (BROEKSEMA; TELEA; BAUDEL, 2013) describes how an MCA (Multiple Correspondence Analysis) could be used to analyze relationships, patterns, trends, and outliers among high-dimensional data. Another option is to take the data from the Object file and consider user-object associations. For each dataset it would be:

- Movielens dataset: We consider the number of times a user reviews a particular movie genre and the average ratings a user has for a movie genre.
- BookCrossing dataset: Based on the same case of Movielens, we have the number of times a user reads one genre of book and the average rating for a genre of book.
- Health dataset: We create a breakdown of the medical specialties (e.g., Respiratory, blood) where each dimension could be interpreted as the number of times a patient was treated.

During the elaboration of this work, we performed different types of experiments on which data to be used as dimensions. In Figure 3.3, we can see a test for the data set Movielens using dimensionality reduction with t-SNE technique. This technique allows to reduce the data space from high dimension to low dimension, and it will be explained in more detail in the next section.

In Figure3.3.(a) we applied the t-SNE algorithm to reduce the 40 dimensions(gender, age, the occupation of each user and genre of movies liked by users) for the Movielens. In this test, we colored the results by gender (Male and Female) of each user. We observe a clear separation of the user by gender. This reflects the great influence of gender in the projection since gender is a categorical data. Now, in Figure3.3.(b), (c) and (d),

Figure 3.3: Dimensionality reduction test for Movielens dataset in different approaches.

(a) All Dimensions (b) Dimension Gender (c) Dimension Age (d) Dimension Occupation

we show the projection of data of Movielens using just the dimensions Gender, Age and Occupation respectively. We can get an intuition on how the algorithm works with categorical data. Thus, we concluded that using categorical data in the projection algorithm is complicated and depends on the type of data. We also concluded that if we are going to use categorical data, all dimensions should be just categorical data and not mixed with other types of data. Therefore, we decided to use quantifiable data as dimensions. For our datasets, they are the number of movies or books of a genre that the user reviewed or the number of medical consultations for a specialty that a user attended.

3.3 Data Projection to $2D$

In this section, it is explained how to project high-dimensional data into low-dimensional data such as 2D. We use a Multi-Dimensional Projection (MDP) technique called t-SNE, widely used in areas such as machine learning, visual analytics, and data mining. This technique was recognized for its potential uses for different types of multidimensional data. We selected this technique for its exceptional visual results demonstrated in the previous studies(Pezzotti et al., 2017).

Below we give a brief explanation of t-SNE, which is necessary to explain our work. The t-SNE algorithm calculates and interprets all distances between data points in the high-dimensional space as one symmetric distribution of probabilities P. In the same way, another probability distribution Q is calculated for low-dimensional data. The purpose of generating these two distributions is to prove that Q is a faithful representation of the P distribution. It means that the projection obtained in the low dimension should be as similar as possible to the same data in high dimension.

This process is achieved by optimizing the positions in the low dimensional space to minimize the cost function C from Kullback-Leibler (KL) divergence between the distributions P and Q.

$$C(P,Q) = KL(P \parallel Q) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} P_{i,j} \ln \left(\frac{p_{i,j}}{q_{i,j}} \right) \tag{3.1}$$

Given two data-points x_i and x_j in the dataset $X = \{x_1..x_N\}$, to model the similarity of these points in the high-dimensional space, we use the representation P_{ij}. Also a Gaussian Kernel P_i for each data-point and one variance σ_i which defines the local density in the multi-dimensional space.

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}, \tag{3.2}$$

$$where \quad p_{j|i} = \frac{exp(\frac{-(\|x_i - x_j\|^2)}{2\sigma_i^2})}{\sum_{k\neq i}^{N} exp(\frac{-(\|x_i - x_k\|^2)}{(2\sigma_i^2)})} \tag{3.3}$$

In the equation 3.2, $p_{j|i}$ represents a measure related to the similarity based on the distance of a neighborhood from a point x_i. This derives from another variable called perplexity (μ). It is a parameter that is used to describe the number of neighbors for each point in the dataset. Perplexity is a variable that can be modified depending on the dataset that is being used. Value σ_i is selected such that for a fixed μ in each i corresponds to the following equation:

$$\mu = 2^{-\sum_j^N p_{j|i} log_2 p_{j|i}} \tag{3.4}$$

The t-SNE technique comes from SNE modified to use a t-Distribution with a degree of freedom. This distribution is used to calculate the probability distribution in a low dimensional space Q, where the points should be optimized to contrast with the high dimension space. The similarity in a low dimensional space can be written given two points y_i and y_j, following the following formula:

$$q_{ij} = ((1 + \|y_i - y_j\|^2)Z)^{-1} \tag{3.5}$$

$$with \quad Z = \sum_{k=1}^{N} \sum_{l=k}^{N} (1 + \|y_k - y_l\|^2)^{-1} \tag{3.6}$$

During the optimization that minimizes C(see Eq. 3.1), the Kullback-Leibler divergence between P and Q is used. It also indicates the change of position of the points in low dimension by each iteration of the descendant gradient, which is given by:

$$\frac{\delta C}{\delta y_i} = 4\sum_{i=1}^{N}(F_i^{attr} - F_i^{rep}) \tag{3.7}$$

$$= 4\sum_{i=1}^{N}(\sum_{j\neq i}^{N} p_{ij}q_{ij}Z(y_i - y_j) - \sum_{j\neq i}^{N} q_{ij}^2 Z(y_i - y_j)) \tag{3.8}$$

According to the author, the gradient can be seen as a comparison of N-bodies, where each point applies a force of attraction and one of repulsion against all other points. These forces are represented in the equation as F_i^{attr} and F_i^{rep}.

4 VISUALIZATION DESIGNS

The main focus for the group exploration process in VUGA is the visualization and interaction with user data. The interface is composed of different areas that show the information of users and other types of actions such as selection, filter, and user saving. Also, it is possible to change configurations for the generation algorithm of user groups. In Fig. 4.1 we have a comprehensive view of the entire system VUGA. There we can distinguish all the components.

The first component is the projection area in Fig. 4.1(a), where we can observe a collection of points and where each point represents a user. The second component Fig. 4.1(b) is called user attribute area, and it is a collection of visualizations for the attributes, the amount of graphics shown in this component depends on the amount of demographic data, this component is customizable by the user in the visual preprocessing stage. In Fig. 4.1(c, d) there is a list with user interests and another list with user information respectively. While in Fig. 4.1(e, f) we have a component called Save Area, which serves to save users of interest and customize the algorithm for generating user groups. In Fig. 4.1(g) the result of the group generation algorithm is shown. Below, each of these components will be explained in more detail.

Figure 4.1: System Overview of VUGA

4.1 Projection Area

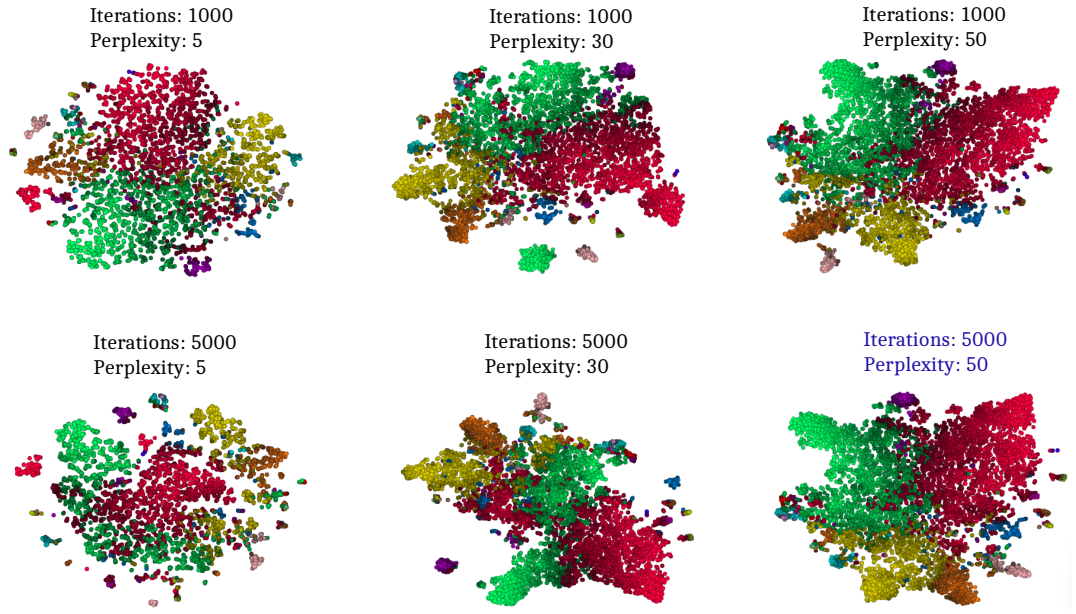
This component is the starting point of the user within the system; here all available data is shown to be analyzed and explored. This component provides a generalized vision about the behavior of the data and it proposes us in advance ideas of things that we might want to discover.

This visualization is the result of applying $t-$ SNE to the high dimension data towards a 2-dimensional space. In Fig. 4.2 we can observe a set of tests performed on the Movielens dataset using two t-SNE parameters: The number of iterations, which indicates the maximum number of times that would be optimized until a solution can be found optimal and adequate. The second parameter is the Perplexity. Perplexity, as previously explained, means a distance from a point to its other neighboring points. The reason why this combination of parameters was used to generate the Projection Area is given by Van Der (MAATEN; HINTON, Nov 2008), where according to his work of $t-$SNE, this is an algorithm that works differently for each dataset. That is to say that depending on the data we would have to experiment $t-$SNE in different situations until finding acceptable visual results for the user's purposes. The author tells us that the recommended values to customize the iteration number are 1000 and 5000 iterations according to the experiments done in previous works. While for perplexity the recommended amounts are 5, 30 and 50.

According to Fig. 4.2, for Movielens dataset, if we use 1000 iterations with a perplexity of 5, the result is a projection of the data with much overlap and with a non-data distribution very clear. This projection could hinder an easy distinction between all available data. With the same perplexity (5) but with 5000 iterations, we have a slightly improved distribution, but still, the overlap of users persists. With a perplexity of 30 and 1000 iterations, it can be seen that it improves a lot, but a proper distribution of the data in the $2-$D space is not achieved, the same happens with 5000 iterations, in which improves slightly compared to the previous one.

However, with a perplexity of 50 and 1000 iterations, we can see that the data managed to differentiate with great clarity but still do not reach a good visual distribution. This projection is finally achieved with a perplexity of 50 and 5000 iterations. In this case, a clear visual distribution and a good differentiation of the data were achieved. Through observation, we could select the final projection, but let us see a variable over $t-$SNE. The C variable explained in the equation 3.7. This variable is minimized during

Figure 4.2: $t-$SNE applied to Movielens dataset using a combination of two parameters: Number of iterations and Perplexity.



the optimization. $t-$SNE algorithm reduces the distances of the nearest neighbors that is optimized in each iteration by a global projection. In the table 4.1 we have the variable C for Movielens dataset using the combinations of Perplexity: 5, 30 and 50 with 1000 and 5000 iterations. Each combination was executed three times to see that there was much variation in each result. C is shown in each combination and where the lowest value is found with a Perplexity of 50 com 5000 iterations that visually was also our selection.

To select a suitable projection using $t-$SNE does not depend on the perplexity or number of iterations or C minimized. These parameters are referential and help us understand our data. An example of this is the BookCrossing data. In Fig. 4.3, we can observe the different projections obtained using the two parameters already described. In this dataset, the projection was chosen with a perplexity of 30 in 1000 iterations. Additionally, we can see that in Table 4.1 it was not necessarily the projection with the lowest C between all the combinations of the parameters. However, that projection that was best visually distributed in two-dimensional space was overlooked.

Additionally, we have the third dataset (Health dataset) used in this work. This dataset has a behavior very different from the first dataset (Movielens). As shown in the figure 4.4 and in the table 4.1. The results of $t-$SNE for this dataset are shown as user groups well separated from others, and there is little similarity among all users. Even with the combination of parameters of $t-$SNE, the projection is not visually pleasing, but on the other hand, it shows some conclusions at first glance, such as the most visited medical

Figure 4.3: $t-$SNE applied to BookCrossing dataset using Number of Iterations and Perplexity as Parameters.

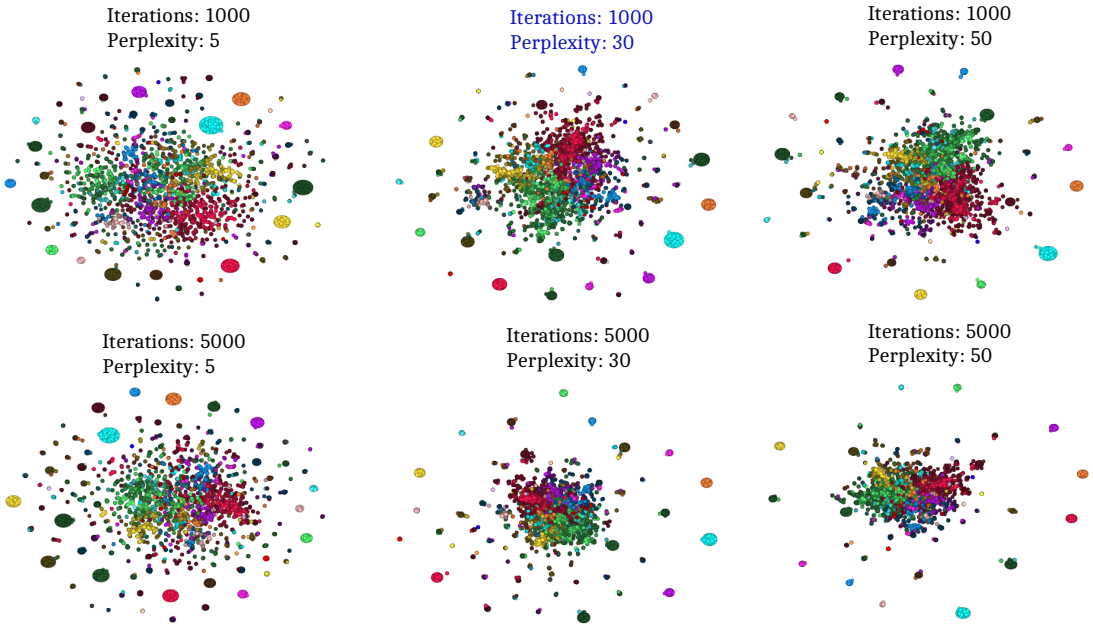


Figure 4.4: $t-$SNE applied to Health dataset using a combination of two parameters: Number of iterations and Perplexity.

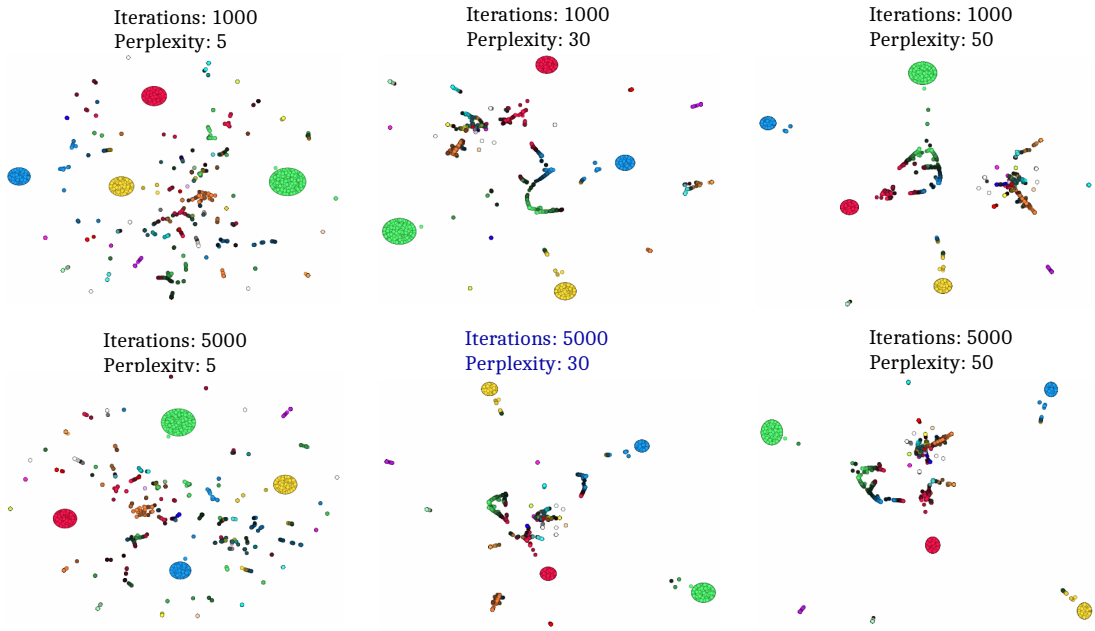


Table 4.1: Movielens, BookCrossing and Health dataset with Parameters: Optimization of the Permutation of Number of iteration with Perplexity executed three times.

Configuration			Optimization in C		
# Run	# Iterations	Perplexity	Moviens	BookCrossing	Health
1	1000	5	1.802657	1.199160	0.887237
2	1000	5	1.765758	1.203062	0.883882
3	1000	5	1.778850	1.193867	0.887607
1	5000	5	1.575450	1.026430	0.860260
2	5000	5	1.556541	1.027299	0.859454
3	5000	5	1.613056	1.025825	0.854147
1	1000	30	1.693772	1.122926	0.657508
2	1000	30	1.705552	1.134787	0.666628
3	1000	30	1.673824	1.135139	0.659744
1	5000	30	1.649885	1.097255	0.649984
2	5000	30	1.672916	1.095400	0.654635
3	5000	30	1.669642	1.087319	0.649512
1	1000	50	1.570772	1.070013	0.593591
2	1000	50	1.578098	1.077052	0.589444
3	1000	50	1.5696902	1.074400	0.588832
1	5000	50	1.5583727	1.051515	0.582760
2	5000	50	1.5557487	1.040932	0.581945
3	5000	50	1.5586364	1.043676	0.585248

specialties.

The use of these three datasets with $t-$SNE shows us the different behaviors that could be obtained with this algorithm. Movielens dataset shows us that their data are much more related among all of them, but they also show a clear separation of data based on their dimensions (like for example the genres of movies). On the other hand, Health dataset shows us a more extreme projection where many of the patients are entirely separated by not sharing similarity with other patients. Few patients had varied similarity with all dimensions. However, the projection for BookCrossing dataset was more conservative, having results similar to those of Movielens and Health dataset. We can observe that some dimensions (genres of books) are exclusive to some people, while others are for the public in general.

This area has some components which help to interact with data. In Fig. 4.5 we can see the area projection and their components. There, the main component is the projection of the users, we create a nD point for each user, containing at each dimension the percentage of interest for each characteristic that the user was interested, for instance in the Movielens dataset each dimension has the percentage of reviews for each genre of movies that the user reviewed. So it is projected this nD space to 2D space using $t-$SNE

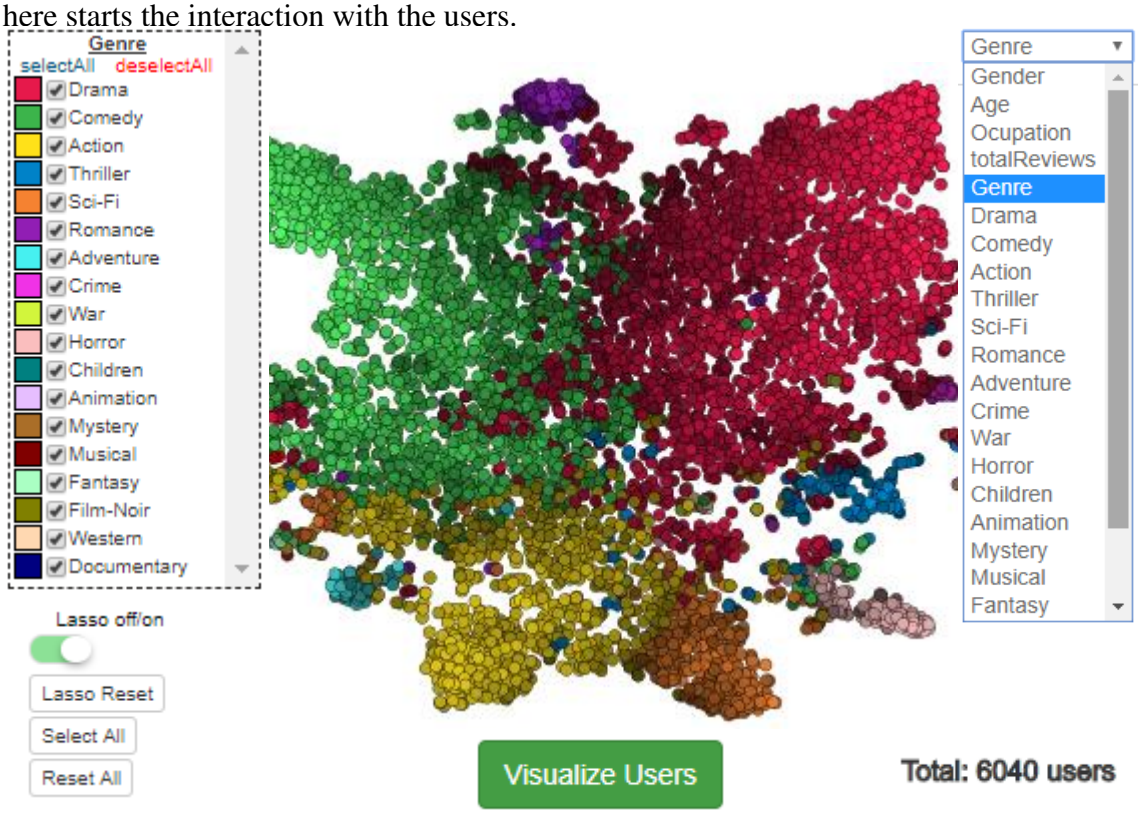
as it was explained previously, and displayed in the area of projection. Additionally, in 2D space, we color each point(User) by the dominating dimension(for instance: users with most reviews of drama genre) following the color scheme shown in the Fig. 4.5 Apply these colors in the points of the projection allows seeing how the projection technique works, creating an environment that minimizes the distance between two or more points(users) with similar dimensions.

The brightness is an additional attribute in the visual representation of the projection. This brightness represents the importance of a dimension in a user. We use more brightness when the dimension(represented as a color) is more important for the user and less brightness when the user has more than one important dimension. Additionally, other components are shown in the area projection. The color selector is a component that contains a list with all the dimensions, the features of users and one selector which shows all dimensions at the same time(for instance, the genre of books in BookCrossing dataset), this color selector allows to select one of the options to color the points(users), for instance in Fig. 4.5, we selected the Genre of movies that are also the dimensions in this dataset, the result was color the users with dominant dimension by each user. The component Legend complements the projection area and is used to see the actual scale of color according to the coloration selected in the color selector component. Additionally, it shows the color associated with each label in legend, for making more interactive, this legend has one check-box by each color available that allows filter and display only the users with the option selected in the projection of the users, so we can separate the users by coloration(feature) and make filters more detailed. In VUGA, it is possible to select one or more subsets of users in the projection using a Lasso Tool Component, this lasso tool works with the mouse and allows a custom control to the user. After, to select one or more subset of users, we can use the ”Visualize Users” button for exploring the details of the users selected. These functionalities are explained in the next section.

4.2 Visualization of Details

As part of the system VUGA, the details of each dataset must be visualized in order to improve the understanding of the meanings of the data and to be able to develop an excellent visual analysis for further steps. After selecting a subset of elements in the previous step, a set of visualizations is created, these visualizations are important to help us to understand the data and see what characteristics it contains. For explaining the

Figure 4.5: Projection Area of Movielens dataset, the first component of visualization, here starts the interaction with the users.



different visualizations for details in this work, we use the Movielens dataset. The analyst will be able to explore in detail the selected users using these visualizations. It will be able to make filters, select some of them, make visual queries through the charts. Then, we aim that the analyst can understand the subset selected from the projection area and can make subsequent decisions based on the detailed analysis offered by the tool.

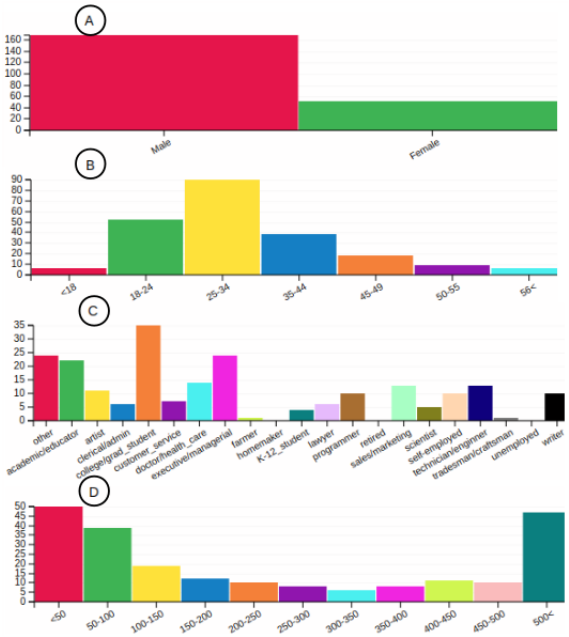
This set of visualizations for the exploration of details can be displayed using different type of visual graphics as: Tables, Line Charts, Pie Charts, Matrices, Histograms, Brush Chart. Each type of characteristic can be shown in some of these graphics depending on its domain. For example, we can use the pie bar to show the gender(male and female).

The interaction between all these visualizations is done under the Crossfilter library(INC., 2012). This library allows joining several graphics under the same query engine, allowing the interaction of a graphic to modify other graphics in real time with high efficiency.

The first visualizations for the elements selected in the previous step are a set of visualization that represent the frequency of users by each attribute of the user. Using

Movielens, we have four visualization charts, for Age, Gender, Occupation and number total of reviews. So, each visualization chart shows a distribution of frequency of users by each label of an attribute of the user. For instance, for Fig. 4.6, we selected 210 users from the area of projection, and these visualization charts were displayed. In part A from the figure, we have the chart correspondent to Gender, this chart shows us that 160 of the users selected are male while that 50 of the users are female. In figure part B, the chart correspondent to Age shows us that the users with age between 25 and 34 are the most dominant of 220 users selected. In the chart of Occupation, the main occupation was college/grad students. The last chart is the chart for Number total of reviews by a user, where most people made more than 500 reviews.

Figure 4.6: Visualization for details of Movielens dataset, A) Gender, B) Age, C) Occupation, D) the Total number of reviews.



The cross-filter will be used to develop these visualization charts. Cross-filter connects all visualization for details of this application. This library allows us to filter the data currently showed, clicking in one bar of any charts, a query is sent to cross-filter, and this will return us a subset of data, updating all visualization charts and even updating other visualization as projection area and tables that are explained below. Cross-filter was also very useful in term of code because this tool has enough support to implement our set of dimensions(layers in cross-filter) that allow getting the responses in a time very short.

The next visualization is a table for passive objects(Movies in Movielens), this table depends on the data filtered by interaction at any moment of the exploration. This table shows the movies that users watched and reviewed. As the table in Fig. 4.7(b) shows

us, the attributes of the movies are the columns of this table. We can order the columns and see the data of different orders. Moreover besides the attributes of a passive object; we show two additional columns, the first is a quantity that means the average of rating received from the users currently filtered on the interaction of the application. The second column is the number of reviews that the passive objects received, this number ever will be less than the number of users filtered in this moment of the interaction.

Figure 4.7: Table for movies of Movielens dataset.

(a) Users

ID	name	Gender	Age	Ocupation	totalReviews
1	user966	M	56<	artist	150-200
2	user957	M	35-44	academic/educator	300-350
3	user94	M	25-34	technician/enginner	<50
4	user884	M	<18	K-12_student	50-100
5	user874	M	45-49	scientist	<50
6	user873	M	56<	doctor/health_care	150-200
7	user868	M	50-55	technician/enginner	50-100
8	user860	M	18-24	other	<50
9	user859	M	25-34	programmer	150-200
10	user808	M	25-34	executive/managerial	250-300

Showing 1 to 10 of 256 entries Previous 1 2 3 4 5 ... 26 Next

(b) Movies

ID	Title	Rating	Genres	# Reviews
1	American Beauty (1999)	4.39	Comedy|Drama	133
2	Silence of the Lambs-The (1991)	4.4	Drama|Thriller	112
3	Schindler's List (1993)	4.61	Drama|War	109
4	Fargo (1996)	4.29	Crime|Drama|Thriller	107
5	Saving Private Ryan (1998)	4.34	Action|Drama|War	105
6	Shawshank Redemption-The (1994)	4.43	Drama	100
7	Braveheart (1995)	4.12	Action|Drama|War	86
8	One Flew Over the Cuckoo's Nest (1975)	4.56	Drama	86
9	Godfather-The (1972)	4.45	Action|Crime|Drama	85
10	Pulp Fiction (1994)	4.31	Crime|Drama	83

Showing 1 to 10 of 2,776 entries Previous 1 2 3 4 5 ... 278 Next

Fig. 4.7(a) shows the data corresponding to the users where the necessary information is showed in the same way that the previous table in Fig. 4.7, for this table with Movielens dataset, the attributes of the users are Name, Gender, Age, Occupation and Number total of reviews. This table allows to order the data from each one of these columns and also use the pagination for having a better order of the elements. Each one of these attributes has its visual representation as a visual chart shown in Figure 4.6, and this table also is updated each time we change the selection of users in the projection area and each time that we filter a sub data using the charts with Crossfilter. Besides that this table allows selecting users clicking on one or more rows of the table, after selecting them they can be saved in another container called Save Area.

Both tables try to improve the experience of the user. This tables search for specific data within their options of filters and show a range of results by each page, these characteristics described improve the user experience about data exploration. Save Area is a region of this application to save the users selected in the table for users, the goal with this area is that we can do a better evaluation of our users selected. When the users are placed in the Save area, the projection area is updated with just the users moved to the Save area. Besides that in this area it is possible to save a file with the users selected, this allows to have a database with lists of users that could have been interesting for the

person that uses the prototype.

The next step in the interaction of the visualization tool is to generate k groups, where k is a number that the user defines. These groups contain a set of users. The importance of these groups is that each one of them has a set of users that have a certain similarity with the original group(set of users saved in the Save area component), this similarity can be: very similar or very different. Furthermore, for example with groups very similar, the users in one group could be different from other groups but similar to the original group. This concept means that a group could have a subset of characteristics similar to the original group and another group could have the complement of characteristics similar to the original group. Finally, all the group could be very similar. It is common that people like to search for similar things, but there are moments when we also want to look for interesting results without considering the preferences of the analysts. It is for that, we also propose the idea of search the opposite thing, with a menu in the prototype it is possible to indicate that we want to search the similar groups and the different groups. In the next section, we describe the generation of new groups based on an original group.

5 DISCOVERING NEW USER GROUPS

5.1 Algorithm for discovering new user groups

In the first part of our proposal, we developed the projection of users from a high-dimensional space to low-dimensional space. We got to establish a point of start to explore the data with a set of visualization that helps to understand the data. In the next step, we propose a way to discover new user groups.

We introduce the idea that discovering new groups of users from a customized group could bring better results. In (OMIDVAR-TEHRANI; AMER-YAHIA; TERMIER, 2015) it is explained that the primary application of data-driven research is the analysis of user data. They also introduce a novel algorithm for discovering user groups. For our purposes, we decided to implement our algorithm for discovering new groups because we need interaction in real time. Thus, we present a new algorithm for discovering a new group of users and a new visualization in VUGA.

In algorithm 5.1, we describe the process for discovering new Groups. This algorithm is a greedy version, but our idea was to implement the first version of our proposal and see how to work the interaction and exploration of users groups. The implementation was developed in the language Python so that the code was optimized in several parts using properties of the programming language.

The input of this algorithm is $dataset, Group0, N$ and $similarity$. The parameter $dataset$ represents a dataset(Movielens, BookCrossing or Health) their dimensions included, $GroupO$ is a list with the users selected and saved in the Save area, we consider that this set of elements is the Original Group or Group Zero. The next variable N is a number which means the number of new groups that the analyst wants to discover. This variable also is specified in the visual interface. By default, this variable is equal to 5 because it is a not so small and not so high amount. Also, helping the interaction does not become massive, but it is the decision of each analyst. The last parameter is the similarity, this variable says the algorithm if we want the most similar or the most different groups, we use MORE when we want the more similar groups, and we use LESS in the opposite case.

The algorithm is divided into two parts, the first is to find all the neighbors for each user according to the type of similarity and the second part is to generate the new user's groups using neighbors found in the first part of the algorithm. This algorithm has

Algorithm 1 Algorithm to Discover new User Groups

Input: dataset, GroupO, N, similarity in
Output: newGroups out
  $ListNeighbors \leftarrow []$
  for $user$ in $GroupO$ **do**
    if $(similarity = MORE)$ **then**
      $ListNeighbors \leftarrow ListNeighbors \cup GetListMoreNeighbors(user)$
    else
      $ListNeighbors \leftarrow ListNeighbors \cup GetListLessNeighbors(user)$
    end if
  end for
  $newGroups \leftarrow []$
  for $n \leftarrow 1$ to N **do**
    for $list$ in $ListNeighbors$ **do**
      $newGroups[n] \leftarrow newGroups[n] \cup GetRandomUsers(list)$
    end for
  end for
  return $newGroups$

a complexity of $O(n^2)$.

In the first part of the algorithm, a local variable $listNeighbors$ is created, which will store all the neighboring users related to each user of $GroupO$. We iterate for each user of $GroupO$, where it is checked if the similarity is $MORE$ (greater similarity) or $LESS$ (less similarity). If there is a search for users with greater similarity, then we look for the users more similar to the current user in $GroupO$ of the loop using the function $GetListMoreNeighbors$. Otherwise, we calculate the farther neighbors using the function of $GetListLessNeighbors$.

These functions use the algorithm KNN(K-Nearest Neighbors) return a list with the users more similar or more distant to a specific user of the original group. In the process of creating this list, it is necessary to make a comparison with each user in the $dataset$ and make a ranking from the most similar to the most different. The criteria for selecting the list with the closest or most distant is using k (=50). We determine the value of K as a result of minimizing error in a K-Fold Cross-Validation (BARIGOU, 2016). Also, we use a radius r that is the distance between two users. To calculate r, we use the maximum distance between two users in $dataset$. Based on the uniform sampling(BEYROUTHY; FESQUET; ROLLAND, 2015) theory, we decided that r is the 10% of the larger distance. So, to select the set of closest or distant users we take into account all the users that are within the radius r of searching for a user in $GroupO$. If the number of users exceeds k, then we only consider the first k users of the list.

In the sequence of the algorithm, we generate new groups. In this second part, we created a $newGroups$ list to store the information with the new users. In this part, we have two loops, the first for the number of new groups N and the second for each list of $ListNeighbors$. The idea is that in each iteration you can extract a user randomly from $list$ and assign it to the new group n. In this way, we ensure that each new group will have the same number of users as in the original group. However, it is possible to have fewer users in a group, this occurs when there are not enough neighbors in the given radius r.

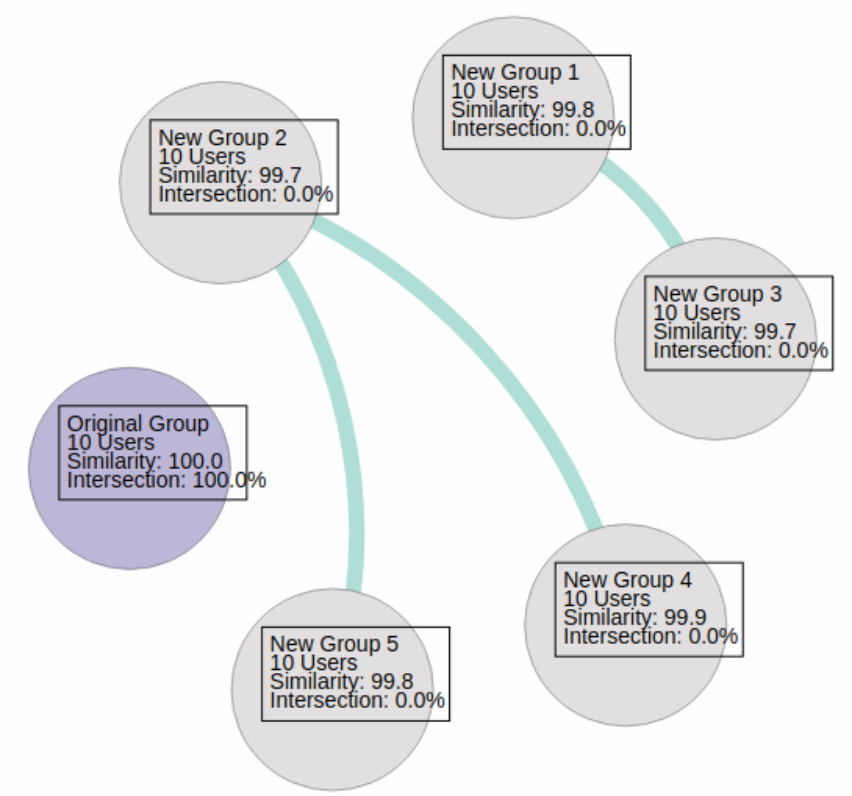
The idea of having the same number of users in all the groups is to be able to make a fair comparison between all of them (original group and new groups). If the user does not want this property, from the graphical interface, it can be customized to say how many users I want in each new group. However, that customization does not modify the main idea of the algorithm. Additionally, another property is included to be able to tell the algorithm what percentage of dimensions I will consider for the generation of new groups. By default it will always be 100%, that is, all the dimensions. If for example, we say that we want 30% of Top Dimensions, then the algorithm will consider only a third of the dimensions by relevance in its distribution. Like the previous custom property, it is a parameter by default of the algorithm and is just an additional property that does not need to be modified by the user.

5.2 Visualizing new users groups

After having generated the new groups, we include new visualizations to display the new groups and the information that these groups have. The new visualizations are to show the information of the users into the new groups and compare two or more new groups at the same time. It can be between two new groups or one new group with the original group. Since the number of users in the groups tends to be the same, then the visualization is more uniform, and the comparison between them is more clear and fairer.

Figure 5.1 shows an example of new user groups using Movielens dataset. Circles represent the new groups. These circles have two visual properties: color and size, the color is used to differentiate between the original group and new groups. Size is proportional to the number of users within a group. In the figure, there are five new groups and the original group is differentiated by the color purple. The original group has ten users, and according to our algorithm, the other groups also have the same number of users. When the new groups are discovered, it is possible to have one or more users in different

Figure 5.1: Visualization of new groups including the original group differentiated with a different color.



groups even users of the original group. This case happens because in some scenarios the user is very close in the multidimensional space. In our visualization, this phenomenon is represented by an edge between two groups. It means that if two groups have an edge that connects them, they have users in common, also the number of users in common. We use the width of the edge to represent the number of users in common in the visualization and being that the maximum and minimum width of the edge depends on the maximum and the minimum number of users on any two groups shown in the visualization including the original group.

$$D_B(p,q) = -ln(BC(p,q))) \tag{5.1}$$

$where$

$$BC(p,q) = \sum_{x \epsilon X} \sqrt{p(x)q(x)} \tag{5.2}$$

Some characteristics included in the visualization are the name, which is used to differentiate them from each other and to know later what group we are exploring. Another characteristic is the number of users in the group, previously explained. Usually, the number of users will be the same in each group, but in some cases, there will not

be enough users for all groups. One case, for instance, is when the users of the original group are very separated from the rest of the users. In this case, the algorithm will not find many neighbors within the established radius, and there will not be enough users for all the groups.

The next feature for groups is the similarity. This characteristic indicates the similarity between a new group and the original group. For calculating this similarity, it was used the Bhattacharyya distance (PATRA et al., 2015), which is used in statistics for measuring the similarity between two discrete or two continuous probability distributions as shown in Eq. 5.1. This distance is also related to the Bhattacharyya coefficient which is a measure of the amount of overlap between the two statistical samples. The distance Bhattacharyya receives two probability distributions. So, we define one probability distribution for each group and after using the distance to see the similarity between the two groups. In the equation 5.3 it is shown the formula to calculate the probability distribution P of a group G, where the group has a size $|G|$(number of users into the group) and a G_j represents the user j for $1 \leq j \leq |G|$, while Dim represents the matrix of users by dimensions used for the projection $t-$SNE previously explained. The equation returns a probability distribution(vector) P with one value for each dimension, each value represents the relevance of the group with each dimension. We obtain a probability distribution that helps us to know what dimensions are more relevant for each group and what dimensions are more important to compare with other groups.

$$P_i = \frac{\sum_{j=1}^{|G|} Dim_{G_j i}}{\sum_{j=1}^{|G|} \sum_{k=1}^{|Dim|} Dim_{G_j k}} \tag{5.3}$$

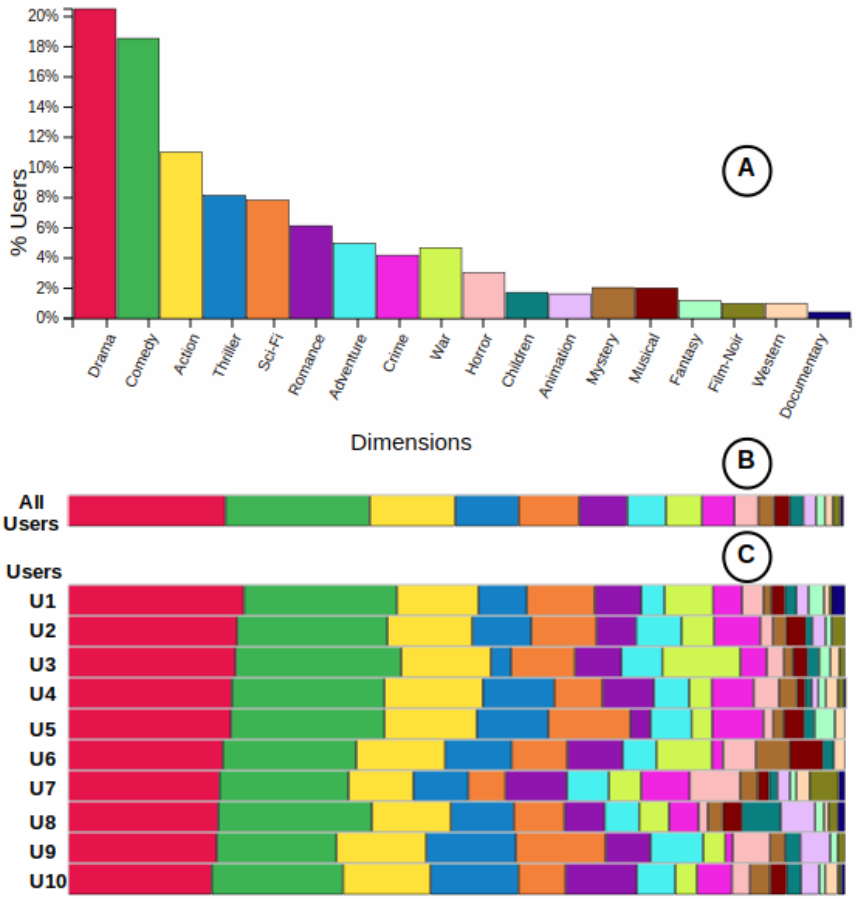
Another feature is the intersection, this feature is displayed in each group and indicates how many users in common have with the original group, as in the case of the group 1 in Fig. 5.1, the intersection with the original group is 0%, meaning that there are not users in common with the original group. So the edge between the two groups is reflected in the variable intersection. So, with the variable intersection, we have a probability of users in common while with the edge we have a visual representation visual and using a tooltip with the number of users in common, which is activated when the mouse is over the tape. However, in the case of the original group, the intersection is always 100%.

5.3 Comparing new users groups

We create three new visualizations to display more statistical information about the new groups and compare two any groups. Fig. 5.2 shows an example using Movielens dataset. In Fig. 5.2 (A), we have a detailed visual representation of the probability distribution of a group. This histogram contains in the horizontal axes the dimensions(18 dimensions in Movielens data) while in the vertical axes we have the percentage of users dominant with each dimension. The colors of these three visualizations are associated with the color used in the projection area. Besides that these colors are as different as possible. In this histogram, for instance, the genre of a movie more relevant for the users is the drama genre followed by the comedy genre.

In Fig. 5.2 (C) we have the probability distribution by each user of Movielens using a type of visualization called stack-Bar Chart. This visualization is sorted by user and use the probability of each gender to compare two users, starting with the genre most relevant obtained in the histogram, if two users are equals in the same genre then we go to the next genre most relevant in the histogram. In Fig. 5.2, the drama genre is the genre most dominant followed by the comedy genre, after action and so on until getting the Documentary genre, which is the less relevant genre for that group of users. While in Fig. 5.2 (B), a stack bar Chart shows the information more compactly, summarizing the information given in the two other visualizations, but with fewer details of the users. The comparison of groups raised here is a proposal of its own that aims to show the similarity between grouped users. Despite the established body of related work to evaluate the exploration of user data, group exploration and visualization only, there is no evaluation methodology for their combination (OMIDVAR-TEHRANI; AMER-YAHIA, 2018; JIANG; RAHMAN; NANDI, 2018).

Figure 5.2: Visualization of the probability distribution of dimensions, a) Histogram with the distribution of preferences of genres, b) Summary of the most relevant genres in descending order, c) Genres importance in descending order by each user.



6 EVALUATION AND DISCUSSION

6.1 Use case example

In this section, we show a use case to show all components of our proposal. This use case is based on Movielens dataset which 6,040 users, 3,952 movies, and 1,000,209 ratings. The interaction begins with the projection area when we select a subset of users according to our preferences. Using the tool, we can have many ways to choose the users; we could select one or more genres using the checkbox in dimension legend or use the Lasso Tool directly.

In Fig. 6.1 we have an example of filters by dimensions, the idea of this filter is to facilitate the interaction and selection of users in the projection area. In part (a) of the figure we have all users of Movielens with all dimensions, but in the (b), (c) and (d) are shown the users for drama, comedy, and horror respectively. For Movielens dataset, the genres most dominants are the drama, comedy and action genre. Now, in this example, we select all user with Horror genre because we want to know more about the users that like the movies of Horror. Selecting all user with Horror genre dominant, we have 118 users of 6040 users in total. After to select these users, we visualize the information in details about those users, so in Fig. 6.2 the ten movies most rated are shown, where we have movies like the Scream(1996), Alien(1979) or Jaws(1975). Some films are not just horror but also have some other associated genre such as the movie Alien, which is also of Action, SciFi and Triller but Horror as the dominant genre of these movies.

The visualization for details is shown in Fig. 6.3 and also is shown how the filters work. In part (a) we have the details of the 118 users selected at the previous step, where 90 users are male, and 28 users are female, most of these people are between 25 and 34 years old, although are users with no occupations register in this dataset, the second occupation most relevant is students followed by executive persons. Also, the majority of the users made from 50 to 100 reviews about the movies.

In this point of the interaction, the filters may be about the features of the users, so part (b) show us an example of filters above the part (a) of the figure. In this part, we select just the persons who are a student; after this filter will update all the visualization in the tool according to the new filter. In this case, now we have just 14 of 118 users, where nine users are male, and five users are female, and where the majority of these users are between 18 and 24 years old, but now many of them made between 100 and 150 reviews.

Figure 6.1: Different types of filters with Movielens using the dimensions. a) All dimension, b) Drama dimension, c) Comedy dimension and d) Horror dimension.

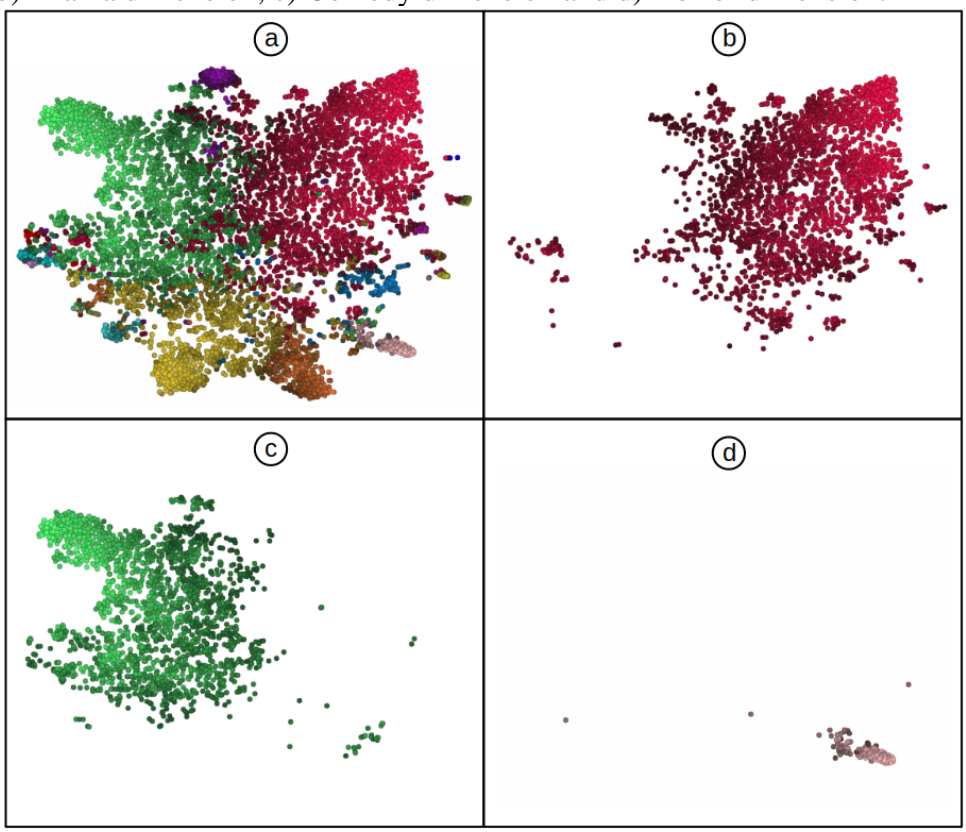


Figure 6.2: Table with the ten movies most rated by users who like Horror movies.

ID	Title	Rating	Genres	# Reviews
1	Scream (1996)	3.83	Horror|Thriller	86
2	Alien (1979)	4.17	Action|Horror|Sci-Fi|Thriller	83
3	Jaws (1975)	4.33	Action|Horror	81
4	Shining-The (1980)	4.26	Horror	76
5	Ghostbusters (1984)	3.85	Comedy|Horror	75
6	Misery (1990)	4.01	Horror	75
7	Psycho (1960)	4.28	Horror|Thriller	74
8	Exorcist-The (1973)	4.31	Horror	72
9	Poltergeist (1982)	3.76	Horror|Thriller	71
10	Carrie (1976)	3.88	Horror	69

Showing 1 to 10 of 1,958 entries Previous 1 2 3 4 5 ... 196 Next

Now we can see, the power of these visualizations is in reducing the number of users to facilitate the exploration of users.

The next step is to choose the users of our interest and create the original group. In this example, we already have 14 users then save these users in the Save area in the

Figure 6.3: Visualization for details in Movielens Data, (a) Before the filter and 118 users, (b) After the filter and 14 users.

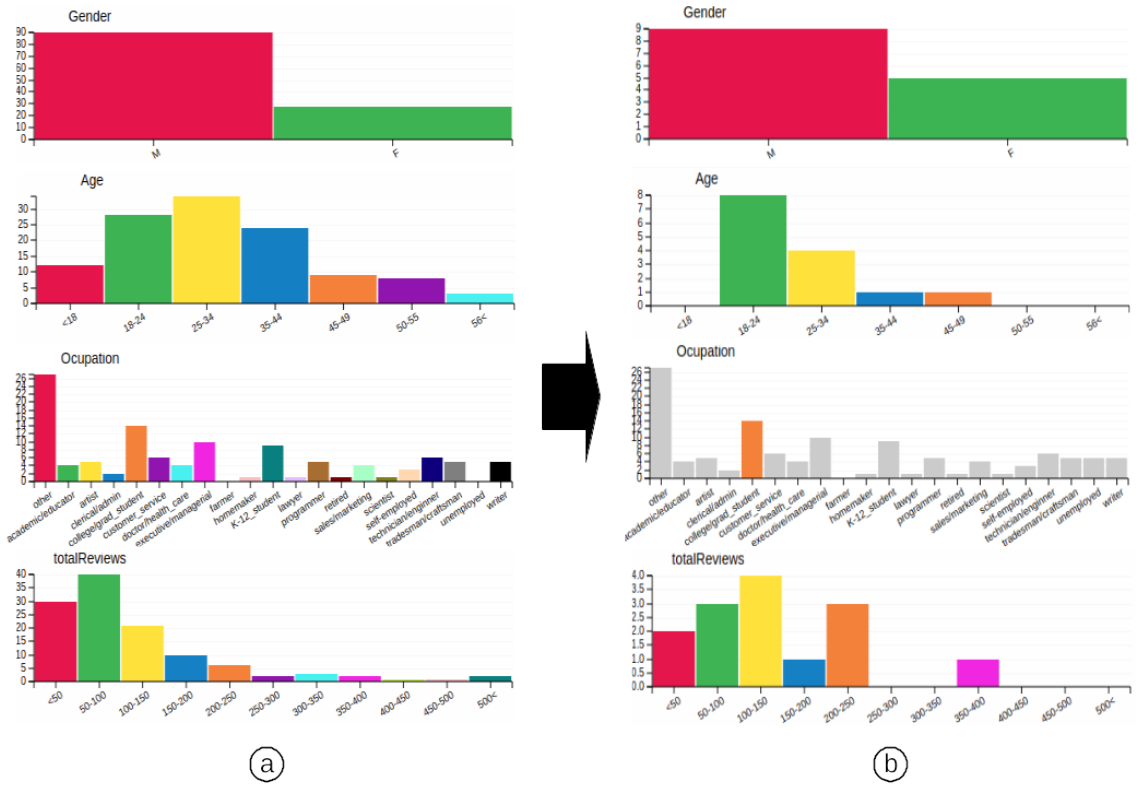
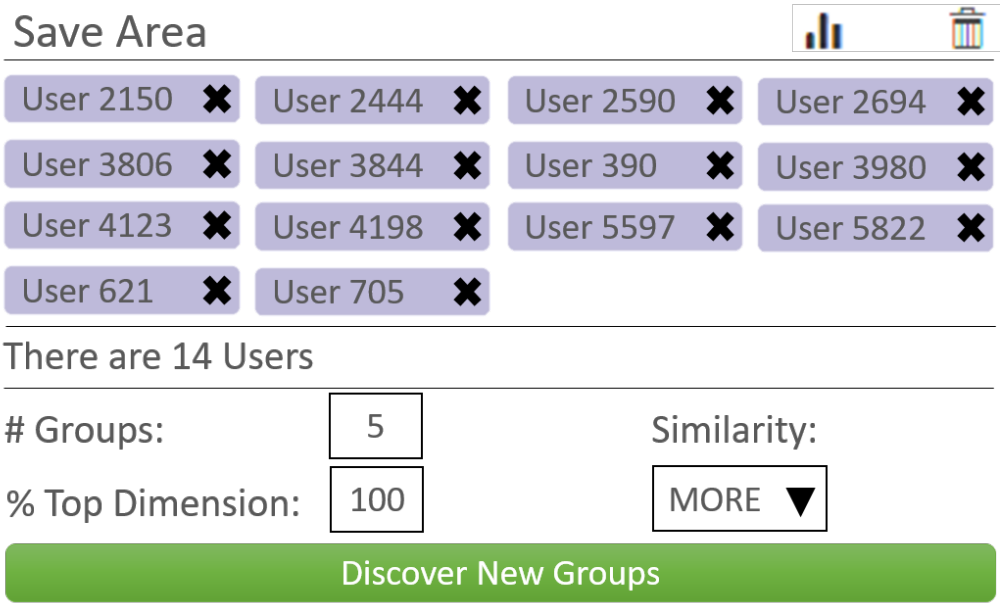


Figure 6.4: Save area for the original group with 14 users and the configuration options for discovering new groups of users.



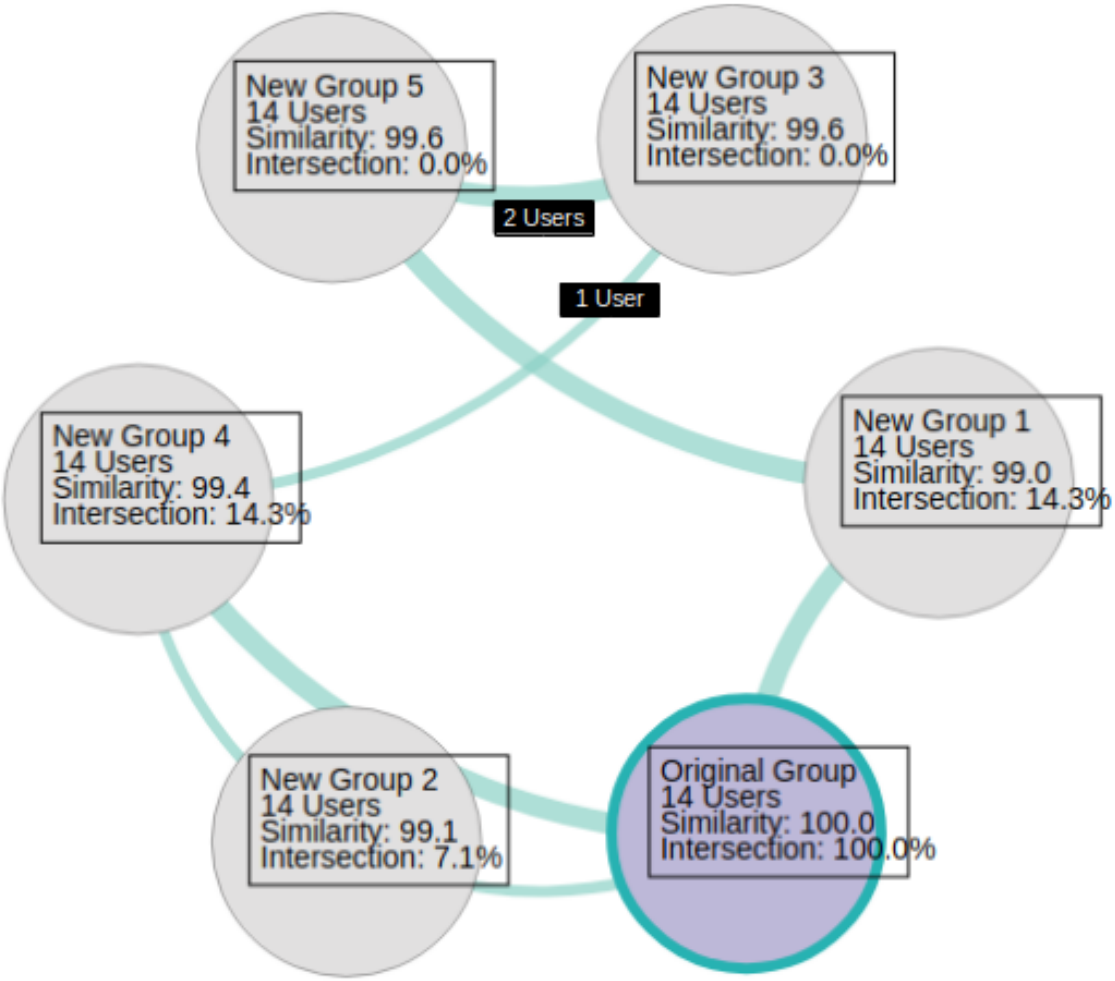
way how shown in Fig. 6.4. In this area, we can remove a user with the icon next to the name of the user or delete all user, and visualize the actual users in the area using the icon on the top. This area was created to prepare the users before to discover new groups.

There are three variables of configuration for generating new groups, the first one is the number of new groups we want, with this configuration we say to the tool that we want n new group with certain similarity to original group, by default we recommend 5 new groups considering that is an appropriate amount for exploring and visualize. The next configuration is the Top Dimensions, that means the percentage of the top most relevant dimension of the original group, for instance. If we would consider the 90% for top dimensions, then this means that for generating the new group we consider the 90%(16 dimensions) most relevant dimensions of 100%(18 dimensions). This configuration is an option to decide how many dimensions we are using to discover new groups, by default, always is in 100% because the idea is to use all dimensions. The last configuration is the Similarity where we have two options, the groups most similar($MORE$) and the groups less similar($LESS$). For this use case, we consider five new groups most similar to the original group using all dimensions(100% top dimensions). After that, we have a button called Explore Groups for executing the algorithm to generate the n new groups.

Clicking on the Explore Groups button the new groups are generated using the algorithm 5.1. In Fig. 6.5 we can see these new groups where there are five new groups with the same color, except the original represented by one different color(purple). There are connections between some groups, and this means there are users in common between the groups with a connection, for instance, the New group 3 have one connection of 1 user in common with the group 4 and other connection of 2 users with the new group 5. Beside each group have a similarity and an intersection with the original group. The original group has a similarity and intersection of 100% because it is the same value as it is a calculation for itself. However the group 1 has a similarity of 99%, this means that is very similar with the original group and also has an intersection of 14.3%, this means that has 2 users in common with the original group.

All groups have 14 users as the original group according to the algorithm which tries to generate new groups with the same size. For seeing the information about a new group, it is necessary to click in them, after this, the information of the users inside the group selected is shown in the visualization for details. Additionally, the visualization for comparing the probability distributions between the original group and a new group selected is activated and displayed in the application. In this use case, the majority of these groups have a high similarity with the original group because the Horror genre is not quite seen in this dataset. Also, there is quite an intersection between groups because the amount of users for this genre is limited. These cases happen because the group of

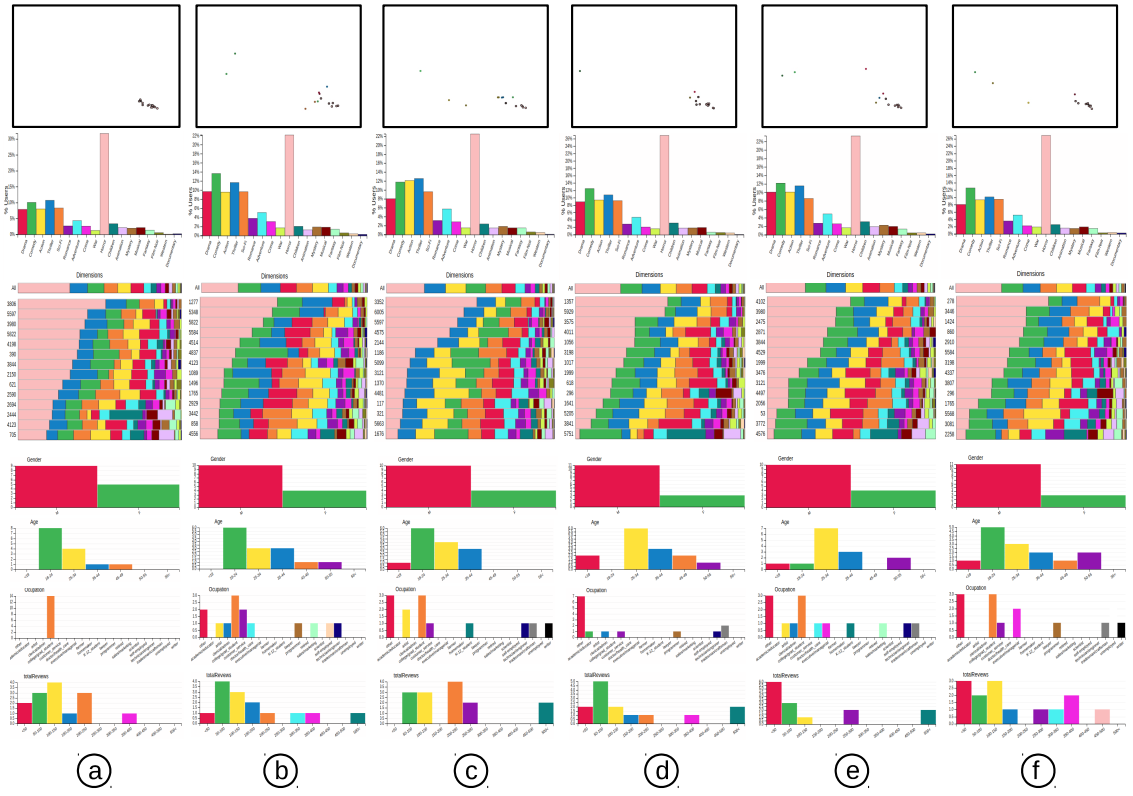
Figure 6.5: Visualization of the original group with the five new groups. Here it is shown the similarity between groups.



users for horror genre is not large, with our approach this could be an advantage because this allows us to explore the users with more similarity as possible and discover the users with high similarity who were not with the group in the projection.

The next step of the analysis is to compare the original group with the new groups. On Fig. 6.6, it is shown the comparison between the original group, which is in (a) and the new groups of users listed from 1 to 5 which are from (b) to (f) respectively in the figure. How we can see in the figure, there is a high similarity between the probability distributions of all the groups. Although there is a similarity among groups, we can find some differences between them. We can see in the stack bar chart summarized of each group, the horror genre is the most dominant genre in the original group and also in the new groups, at least in this example, this genre is most relevant in comparison with the other genres of movies. In the next genres according to the order of relevance change for each group. Remembering that the stack bar chart summarized helps to see the order

Figure 6.6: Comparison between the original group and all new group using the probability distribution summarized and the detailed.



of the dimensions by the genres most dominant. Thriller genre was the second most dominant genre in the original group and group 2(c). These groups make sense for the analysis because Thriller and Horror are similar genres and it is most probable that users being similar. However in the rest of the groups, Comedy genre is the second genre most dominant, this generates interest in knowing that is happening and why Comedy is in this position. Although the third genre most dominant is the Triller and the forth is Action, which is more close to Horror. In this point, we have to see the visualization for details of the users in each group to analyze the groups; at the top of each group in Fig. 6.6, we have one way of seeing what is happening with the users of each group. Reviewing, the original group was formed from users who like just the horror movies, in the figure we can see these users grouped in the lower right corner. How is normal, some users are close to the users of the original group. However, we can see some users of the new groups are further away from the users of the original group. These users also are attractive to analyze them. Additionally, we can see these users have another color different to Horror color, this means these have more preference for another genre of movie but with high similarity to the original group in others dimensions and features.

In the visualization for details of the new groups, we have four charts, for gender, Age, Occupation and total Reviews of the users. In the gender chart, all groups are quite similar. In the Age chart, we have more differences, the original group, and the new groups 1(b), 2(c) and 5(f) are similar because they have the users with age between 18 to 24 years old. However, group 3(d) and 4(e) have ages between 25 and 34 years old. Additionally, we can see that while in the group 1, 2 and five the ages most frequents are of 18 to 24 years old, in the groups 3 and four this interval of ages almost does not exist. For the Occupation chart, we have the original group with just student, the users of the new groups also are students in their majority except in group 3(d). However other occupations are very frequent. In the visualization for the number of reviews, the difference between the group was not very remarkable. As we discovered groups with high similarity, we could also discover groups of similarly low users, that is, groups different from the original group. When we explored groups more different from the original group, we got groups with the high taste for comedy movies along with drama.

In this use case, we explore a set of users and discover other users with similarities. We use users with the same dominant genre because we wanted to show how it is possible to discover users with other dominant genres in their preferences but with similarities in their other genres. In a use case more extended we could select users with different preferences in many genres and with more filters.

6.2 User Evaluation

In this section, it is shown the user study about VUGA. This study shows how useful is our approach and how interactive can be with real users. We design a questionnaire with tasks that test VUGA. Additionally, we used SUS(System User Usability) to evaluate the usability of our approach.

The prototype has a web-based interface which was distributed to subjects. During the study, there was no direct contact with subjects. The first part of the experiments has a tour into the system to explain the user interface and support tasks. For each question formulated, we assessed the difficulty of the subject to complete the task using a Likert scale, from one (hard) to 5 (easy). We formulated three hypotheses to evaluate the results of the study:

- **H1**: Interaction in the projection area allows us to explore and analyze high-dimensional

user data.

- **H2**: The visual analytics interface with filtering capabilities coupled with a group generation algorithm allows performing visual user group analysis.
- **H3**: The visual comparison of groups using stacked-bar charts allows the analysis of similarity in groups.

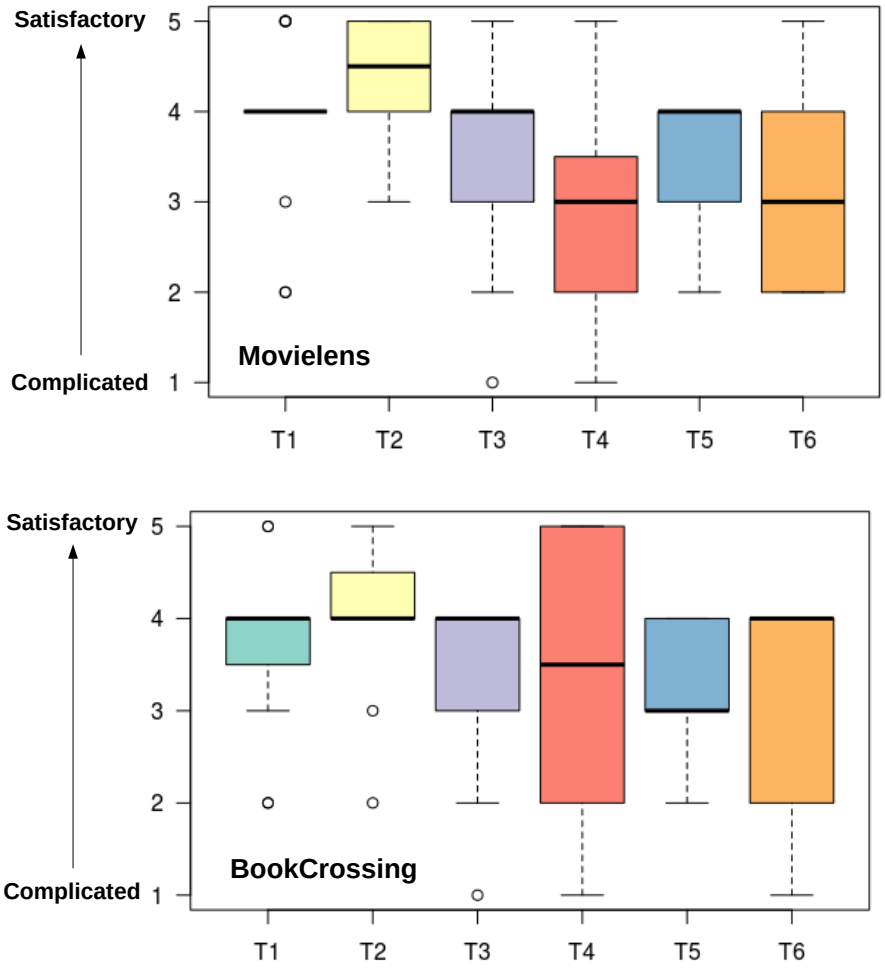
This user evaluation is worked using the Movielens and BookCrossing datasets. This two dataset was chosen because they are similar struct but different aims. We distributed a questionnaire along tasks to 24 subjects, in their majority male (70,83%). The subjects age average is 30 years old (from 20 to 50 years old). Among subjects, 11 are M.Sc students in Computer Science, 4 have specializations in Information Technology, and the remaining nine are from Humanities and Social Sciences. The profile of experience with data visualization shows that 50% know info-graphics used in websites and newspapers, 66.7% used graphics in their works, 58% used data visualization in their work, and 66.7% used data visualizations in their research. We defined six tasks comprehending some actions such as selection and filtering users, saving users, group exploration and analysis. Since both datasets have user reviews of items (movies or books), we define the same set of tasks for both datasets, just exchanging movies by books accordingly. In the study group, 12 subjects used Movielens and 12 used BookCrossing. The list of tasks include:

- **T1**: Select all users that prefer romance movies/books as dominant dimensions, and generate visualizations of their demographics.
- **T2**: Filter the users selected in the previous task into a smaller group of your interest and save them in the Save Area.
- **T3**: Using the original group the users in the save area, we can configure the parameters of the group generation algorithm and explore similar groups.
- **T4**: Find and describe two movies/books in common with the highest rating among similar groups to the start group.
- **T5**: Genres of movies/books in the stacked bar chart are given in order of dominance. Is there any order in a group different from other groups? If it exists, describe what you observe.
- **T6**: What are the genres of movies/books and the location of people more different from the people who have a great preference for Documentary/History?.

Figure 6.7 shows the boxplot of the assessment of tasks completions by subjects

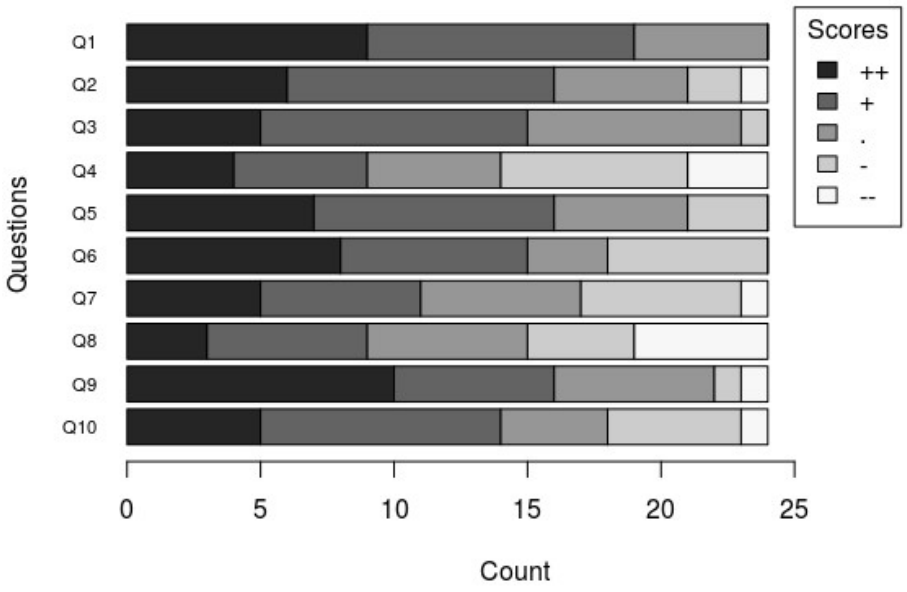
on the Likert scale. We observe that user was successful upon completing tasks T1 and T2, with a median value equal or superior to four. Both tasks are related to exploring and filtering user data in the projected area. They confirm hypothesis H1, which is related to the success of exploring user data in the projection area. In tasks T3-T6 we also asked a text response, which was useful to assess how well the subject understood the tasks to be completed. Most subjects performed the tasks correctly. In task T3, most subjects chose to discover four to five similar groups. In task T4 the subject had to explore groups and find movies/books most common among all groups. This task was the most challenging among all tasks (mean closer to 3), with subjects reporting difficulty in finding similar movies/books. This can be explained by the fact that we design the system around interaction over the user demographics and preferred genres information. The evaluation of tasks T3 and T4 confirms that hypothesis H2 was mostly successful. Task T5 was designed to require visual comparison among user groups. The evaluation was positive for both datasets. This task helped to understand the composition of group dimensions

Figure 6.7: Boxplot of the six task questionnaire and the results using the Likert scale.



concerning the movie/book genre. Finally, in task T6 most subjects also were able to complete the task, but also reported difficulties in using all interactions and visualizations in the interface. We conclude that H3 was partially satisfied and we will use the feedback to improve the design of the user interface.

Figure 6.8: Stack Bar Chart of the six task questionnaire with the results using the Likert scale.



Additionally, we have a questionnaire SUS to validate the usability of the prototype. The questionnaire of SUS is in the appendix. In Fig. 6.8 is shown the distribution for each question in SUS(10 questions). For this figure, we converted the scores of a range of '–'(score 1 on odd-numbered questions odd and score 5 on even-numbered questions), '.' for score 3 to '++'(score 5 and 1 for odd- and even-numbered, respectively). We use a gray-scale to determine the importance of each question, darker color is a positive response for us. The mean score obtained in the SUS was 51,97 with an s.d. of 10.2. The strongest agreement was question 9, "I felt very confident using the system" followed by question 1, "I think that I would like to use this system frequently". Additionally, we asked for personal feedback about the prototype.

6.3 User Feedback

In addition to the task-based assessment and questionnaires explained above, we collected the opinions of the participants in order to improve our proposal for future work. In the following list, we summarize the opinions.

- Projections are a great help to give a quick analysis of what we want to explore.

Generates a general idea of how the data is distributed in the space.

- The system must be a bit more intuitive so that all people can use them without having prior knowledge.
- The concept of generating groups is something very new and interesting to discover new information.
- The response time for the different queries must be improved when new groups are generated in the interaction.

Among the most important opinions, we emphasize that the prototype was interesting to explore for new information and on the other side that should improve the order of some components to enhance the experience of analysis.

7 CONCLUSION

In this work, we described the design of a visualization system to support the exploration and analysis of multidimensional data from user rating products. The central part of our work was to design a visual tool to explore and analyze user groups. We proposed the projection of multidimensional data as a point of start to visualize data in high dimensions. We also proposed an algorithm to discover new user groups from an initial group of users. Finally, we implemented visual components to detail and to compare the new groups. The results obtained revealed interesting insights, where the visualization of the projection of users was well received by the participants as well as the comparison of user groups.

The system has other functionalities that are not explained here, because we consider improving the system in future work. One possibility of future work is to improve the performance of the algorithm to discover new groups. Our algorithm can be improved using spatial data structures such as a Quadtree or kd-tree. Additionally, we want to test the system with additional data sets with more complexity, more dimensions, and more users. We also want to test the VUGA with more analysts and different levels of expertise.

REFERENCES

AMER-YAHIA, S. et al. Exploration of user groups in VEXUS. **CoRR**, abs/1712.03529, 2017. Available from Internet: <http://arxiv.org/abs/1712.03529>.

BARIGOU, F. Improving k-nearest neighbor efficiency for text categorization. **Neural Network World**, v. 26, p. 45–66, 01 2016.

BEYROUTHY, T.; FESQUET, L.; ROLLAND, R. Data sampling and processing: Uniform vs. non-uniform schemes. In: **2015 International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP)**. [S.l.: s.n.], 2015. p. 1–6.

BROEKSEMA, B.; TELEA, A. C.; BAUDEL, T. Visual Analysis of Multi-Dimensional Categorical Data Sets. **Computer Graphics Forum**, The Eurographics Association and Blackwell Publishing Ltd., 2013. ISSN 1467-8659.

Cheng, S. et al. Model-driven visual analytics for big data. In: **2016 New York Scientific Data Summit (NYSDS)**. [S.l.: s.n.], 2016. p. 1–2.

DAS, J. et al. Iterative use of weighted voronoi diagrams to improve scalability in recommender systems. In: CAO, T. et al. (Ed.). **Advances in Knowledge Discovery and Data Mining**. Cham: Springer International Publishing, 2015. p. 605–617. ISBN 978-3-319-18038-0.

DIMITRIADOU, K.; PAPAEMMANOUIL, O.; DIAO, Y. Explore-by-example: An automatic query steering framework for interactive data exploration. In: **Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: ACM, 2014. (SIGMOD '14), p. 517–528. ISBN 978-1-4503-2376-5. Available from Internet: <http://doi.acm.org/10.1145/2588555.2610523>.

DIMITRIADOU, K.; PAPAEMMANOUIL, O.; DIAO, Y. Aide: an active learning-based approach for interactive data exploration. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 28, n. 11, p. 2842–2856, 2016.

DONG, Q.; CHENG, Y.; MIN, Z. Interactive design on recommender system. In: **2017 IEEE 17th International Conference on Communication Technology (ICCT)**. [S.l.: s.n.], 2017. p. 1884–1890.

DUNN JR, W. et al. Exploring and visualizing multidimensional data in translational research platforms. **Briefings in Bioinformatics**, v. 18, n. 6, p. 1044–1056, 2017. Available from Internet: <http://dx.doi.org/10.1093/bib/bbw080>.

FEKETE, J.-D.; PLAISANT, C. Interactive information visualization of a million items. In: IEEE. **Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on**. [S.l.], 2002. p. 117–124.

HU, Z.; YAO, J.; CUI, B. User group oriented temporal dynamics exploration. In: **Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence**. AAAI Press, 2014. (AAAI'14), p. 66–72. Available from Internet: <http://dl.acm.org/citation.cfm?id=2893873.2893885>.

HUANG, M. L.; HUANG, T.-H.; ZHANG, X. A novel virtual node approach for interactive visual analytics of big datasets in parallel coordinates. **Future Generation Computer Systems**, v. 55, p. 510 – 523, 2016. ISSN 0167-739X. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0167739X15000382>.

IDREOS, S.; PAPAEMMANOUIL, O.; CHAUDHURI, S. Overview of data exploration techniques. In: **Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: ACM, 2015. (SIGMOD '15), p. 277–281. ISBN 978-1-4503-2758-9. Available from Internet: <http://doi.acm.org/10.1145/2723372.2731084>.

INC., S. **CrossFilter Crossfilter**. 2012. Available from Internet: <http://square.github.io/crossfilter/>.

ITOH, T. et al. High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots. **Journal of Visual Languages & Computing**, v. 43, p. 1 – 13, 2017. ISSN 1045-926X. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1045926X15300112>.

JIANG, L.; RAHMAN, P.; NANDI, A. Evaluating interactive data systems: Workloads, metrics, and guidelines. In: **Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018**. [S.l.: s.n.], 2018. p. 1637–1644.

JO, J. et al. Swifttuna: Responsive and incremental visual exploration of large-scale multidimensional data. In: **2017 IEEE Pacific Visualization Symposium (PacificVis)**. [S.l.: s.n.], 2017. p. 131–140.

KAMPARS, J.; GRABIS, J. Near real-time big-data processing for data driven applications. In: **2017 International Conference on Big Data Innovations and Applications (Innovate-Data)**. [S.l.: s.n.], 2017. p. 35–42.

KRUIGER, J. F. et al. Multidimensional data exploration by explicitly controlled animation. **Informatics**, v. 4, p. 26, 2017.

MAATEN, L. van der; HINTON, G. Visualizing high-dimensional data using t-sne. **Journal of Machine Learning Research**, v. 9: 2579–2605, Nov 2008.

MAZUMDAR, S.; PETRELLI, D.; CIRAVEGNA, F. Exploring user and system requirements of linked data visualization through a visual dashboard approach. **Semant. web**, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 5, n. 3, p. 203–220, jul. 2014. ISSN 1570-0844. Available from Internet: <http://dl.acm.org/citation.cfm?id=2786122.2786125>.

OMIDVAR-TEHRANI, B.; AMER-YAHIA, S. User group analytics: Discovery, exploration and visualization. In: **Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018**. [S.l.: s.n.], 2018. p. 2307–2308.

OMIDVAR-TEHRANI, B.; AMER-YAHIA, S.; TERMIER, A. Interactive user group analysis. In: **Proceedings of the 24th ACM International on Conference on Information and Knowledge Management**. New York, NY, USA: ACM,

2015. (CIKM '15), p. 403–412. ISBN 978-1-4503-3794-6. Available from Internet: <http://doi.acm.org/10.1145/2806416.2806519>.

PATRA, B. K. et al. A new similarity measure using bhattacharyya coefficient for collaborative filtering in sparse data. **Knowledge-Based Systems**, v. 82, p. 163 – 177, 2015. ISSN 0950-7051. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0950705115000830>.

Pezzotti, N. et al. Approximated and user steerable tsne for progressive visual analytics. **IEEE Transactions on Visualization and Computer Graphics**, v. 23, n. 7, p. 1739–1752, July 2017. ISSN 1077-2626.

POSTGRESQL. **PostgreSQL PostgreSQL**. 2018. Available from Internet: <https://www.postgresql.org/>.

QIAN, D. et al. A visualization and interaction system of multivariate movie network data. In: **2016 IEEE International Conference on Signal and Image Processing (ICSIP)**. [S.l.: s.n.], 2016. p. 379–383.

Rein, P. et al. Group-based behavior adaptation mechanisms in object-oriented systems. **IEEE Software**, v. 34, n. 6, p. 78–82, November 2017. ISSN 0740-7459.

SAKET, B. et al. Visualization by demonstration: An interaction paradigm for visual data exploration. **IEEE Transactions on Visualization and Computer Graphics**, v. 23, n. 1, p. 331–340, Jan 2017. ISSN 1077-2626.

TORNADO. **Tornado Tornado**. 2016. Available from Internet: <https://www.tornadoweb.org/en/stable/>.

Wang, C.; Cao, L.; Chi, C. Formalization and verification of group behavior interactions. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, v. 45, n. 8, p. 1109–1124, Aug 2015. ISSN 2168-2216.

ZANABRIA, G. G.; NONATO, L. G.; GOMEZ-NIETO, E. istar (i*): An interactive star coordinates approach for high-dimensional data exploration. **Computers & Graphics**, v. 60, p. 107 – 118, 2016. ISSN 0097-8493. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0097849316301054>.

APPENDIX A — QUESTIONNAIRE SUS

- I think that I would like to use this system frequently.
- I found the system unnecessarily complex.
- I thought the system was easy to use.
- I think that I would need the support of a technical person to be able to use this system.
- I found the various functions in this system were well integrated.
- I thought there was too much inconsistency in this system.
- I would imagine that most people would learn to use this system very quickly.
- I found the system very cumbersome to use.
- I felt very confident using the system.
- I needed to learn a lot of things before I could get going with this system.