UserDEV: A Mixed-Initiative System for User Group Analytics

Behrooz Omidvar-Tehrani[1], Sihem Amer-Yahia[1], Eric Simon[2], Fabian Colque Zegarra[3], João L. D. Comba[3], Viviane Moreira[3]
[1]Univ. Grenoble Alpes, CNRS, LIG (France), [2]SAP (France), [3]UFRGS (Brazil)
[1]firstname.lastname@univ-grenoble-alpes.fr, [2]eric.simon@sap.com, [3]fbcolque@gmail.com,{comba,viviane}@inf.ufrgs.br

ABSTRACT

The increasing availability of user data constitutes new opportunities in various applications ranging from behavioral analytics to recommendations. A common way of analyzing user data is through "user group analytics" whose purpose is to breakdown users into groups to gain a more focused understanding of their collective behavior. The process consists of group discovery, group exploration, and group visualization. To date, user group analytics is done using separate tools which makes it fragmented and burdensome for analysts. In this paper, we describe UserDEV, a full-fledged user group analytics pipeline which combines discovery, exploration, and visualization of user groups, in a fully-connected fashion. UserDEV contributes a star-like architecture as well as a common data exchange model to tighten connections between the analytics components. We provide a realistic use case to show how UserDEV helps analysts perform analytical tasks on user groups. While we report a preliminary user study, we also discuss opportunities for an end-to-end evaluation of a group analytics framework.

1 INTRODUCTION

Today, user data is ubiquitous in various domains, ranging from the social Web to medical records, scientific publications, and retail store receipts. This data is characterized by a combination of demographics (e.g., age, gender) and actions (e.g., rating a movie, publishing a paper, following a medical treatment). Analysts rely on user data to achieve a variety of tasks with the target goal of finding people of interest or analyzing collective behavior. A common way of analyzing user data is through "user group analytics" whose purpose is to breakdown users into groups to gain a more focused understanding of their collective behavior. Group analytics is a shift towards Quantified-Us [1] and helps analysts make better and faster decisions [2] with more certainty [3]. It also addresses peculiarities of user data such as noise and sparsity. A list of group-centric scenarios in various domains is provided in Table 1.

To obtain insights on user groups, analysts need to go through a group analytics pipeline, which consists of three main components: *discovery*, *exploration*, and *visualization* [4]. Group discovery takes as input raw user data and finds groups that reflect the behavior of a set of users, e.g., "Asian women who publish regularly in databases" [5]. Once groups are discovered, group exploration is used to tackle information overload in the plethora of generated groups by enabling navigation in the group space [6]. To render groups in a human-understandable form, group visualization maps groups to visual variables [7, 8]. The following example describes a realistic scenario of analyzing user groups.

Example. *Emma is an organizer of movie critics panels. She aims to gather a diverse set of reviewers at the first screening of Drama and Comedy movies. For this task, she relies on the MovieLens movie rating dataset.*[1] *She first* ***discovers*** *groups of reviewers with common demographics. In order to pick diverse reviewers for both Drama and Comedy genres, she needs to expand and* ***explore*** *various discovered groups. For a better goal-oriented exploration, she decides to* ***visualize*** *exploration options. Through visual inspection, she handpicks reviewers from groups with different age categories and occupations, and builds a diverse set of reviewers.*

To date, user group analytics is done using separate tools. As these approaches only address isolated parts of the group analytics pipeline, they impede analysts to obtain end-to-end group-centric insights directly from raw user data. In [5, 9, 10], thousands to millions of user groups are discovered using different methods such as subspace clustering and community detection, but it is not clear how analysts can find an interesting subset of results in this voluminous search space. In [7, 11], on the other hand, a visual analytics methodology is proposed to explore user groups as first class citizens, but it is not clear how those groups should be discovered first. In such cases, analysts must connect different analytical tools together (e.g., provide the output of a group discovery or group exploration method as the input for a group visualization method) to manually build a pipeline. The drawbacks of fragmented pipelines are mentioned as follows (**C1** to **C5**).

C1: Lack of connectivity. When group analytics components are not inherently connected (i.e., siloed components), they are agnostic about the functionality of each other, hence the whole pipeline would not function in an optimized and cooperative way. For instance, an analyst wants to visualize two millions of generated groups using a visualization library such as D3[2]. As the visualization layer is not aware of the group discovery's output, it aims to visualize all groups at once, which exceeds browser's buffer limit.

C2: Switching cost. In user group analytics, it is crucial for analysts to be able to switch between group-view and user-view (i.e., members of groups) easily. For instance, an analyst may need to verify users of a generated group in a visual view. Lack of connectivity between components results in switching costs, i.e., the visual component may need to send a new *overhead request* to the discovery component and ask for members of a discovered group.

HILDA'19, July 5, 2019, Netherlands

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

[1] *https://grouplens.org/datasets/movielens/1m/*
[2] *https://d3js.org*

Domain	Group-centric needs
Health-care [12, 13]	How has the health status of a **group of patients** with common symptoms evolved?
Crowdsourcing [14]	What is the most appropriate **group of workers** to perform a collaborative task (e.g., sentence translation)?
Airline [15]	Which **groups of travelers** tend to travel to promoted destinations?
Finance [16]	What are **customer groups** in a bank with common trends (e.g., high attrition rate) for product targeting?
Sport [17]	Which **group of players** has an exceptional performance for improvement planning?
Education [18, 19]	How is the overall progress of **student groups** who are weak in a specific discipline?
Fairness [20]	How much fair different classification algorithms are to **sensitive groups** (identified by race, gender, and age)?

Table 1: Real-world scenarios centered on user groups.

C3: Transfer cost. Inputs and outputs of different group analytics methods do not necessarily match. Hence there should be mediators (adapters) between each pair of components to transfer the output of the precedent component to the input of the ensuing [21]. This impedance mismatch puts burden on analysts to build a pipeline.

C4: Learning cost. Fragmented pipelines lack explainability. As there is no inherent connection between the components, analysts cannot easily learn how an exploration option or a visual view originate from group discovery. Hence there is a learning cost associated to such manually-built pipelines, where analysts require to check data lineage in different components manually.

C5: Lack of genericity. Even if a pipeline is manually built by an analyst to serve her specific group-centric goals, it cannot be re-used for other applications, as it lacks genericity. For instance, C-Explorer [22] provides a visual exploration layer tailored to groups generated by attributed community search, and it needs re-configuration for employing another community detection method.

In this paper, we introduce UserDEV, a full-fledged user group analytics pipeline which combines **D**iscovery, **E**xploration, and **V**isualization of **User** groups, in a fully-connected fashion. UserDEV provides a star-like architecture where both components and analysts are connected to a *central node* (corresponding to **C1** and **C2**). This central node is responsible for handling analyst's input as well as connecting components together, using a *common data exchange model* as a communication means (corresponding to **C3** and **C4**). The common data exchange model captures properties of user data and user groups and is independent from any particular implementation of group discovery, exploration, and visualization (corresponding to **C5**).

While there exists several data analytics pipeline approaches ([23–25], to name a few), the literature on *discover-explore-visualize* pipelines for user groups is sparse, as it is an ongoing research direction. Mixed-initiative systems are proposed in [26–30] to incorporate analyst interactions and enable explorations on raw data sources. However, most visualization insights are in the form of simple histogram and regression recommendations, whereas user groups require more expressivity in visualizations. Picasso [31] and C-Explorer [22] are other instances where user group components are combined to serve one specific aim, i.e., graph and community visual exploration. Lumira [32] and VAS [33] are visual exploration methods which prune discovered groups based on available pixels to visualize. Ermac [34] is a vision to connect different components (in our case, group analytics components) together using a common data exchange model, and execute the whole session with a traditional database engine. All aforementioned approaches suffer from some or all of the drawbacks **C1** to **C5**, as they are tailored to specific applications, and a common communication means is not defined between the components.

Paper organization. Section 2 defines the user data model. Section 3 describes design considerations. Section 4 provides details on the functionality of UserDEV's pipeline. Section 5 reports a realistic use case of user group analytics in UserDEV. In Section 6, we present a preliminary user study which evaluates the usefulness of our framework, and discuss on-going efforts and challenges in evaluating our framework.

2 USER DATA MODEL

UserDEV requires a generic data model to support the variety of user data in different domains. We define user data $\mathcal{D}$ as a quadruple $\langle \mathcal{U}, \mathcal{I}, \mathcal{E}, \mathcal{X} \rangle$, where $\mathcal{U}$ is a set of users, $\mathcal{I}$ is a set of items, $\mathcal{E}$ is set of events, and $\mathcal{X}$ is a similarity matrix. Each user $u \in \mathcal{U}$ is described as $u = \{\langle a, v \rangle\}$, where $a \in \mathcal{A}$ is a demographic attribute (e.g., gender, age, occupation), and v is a value in a's domain, i.e., $v \in dom(a)$. For instance $u_1 = \{\langle gender, female \rangle, \langle age, young \rangle, \langle occupation, student \rangle\}$ represents a young female student. Similarly, each item $i \in \mathcal{I}$ is described using attributes such as movie information or medical treatments. The set $\mathcal{E}$ connects users and items, and contains triples $e = \langle u, i, t \rangle$ which describes that user i performs an action (e.g., rating, watching, voting, consuming) on item i at time t. Last, $\mathcal{X}$ captures similarities between user pairs, and contains tuples $\langle u, u' \rangle \rightarrow [0, 1]$ which associates a similarity value to the pair of users u and u'.

A user group is a triple $g = \langle members, demogs, items \rangle$ where $g.members \subseteq \mathcal{U}$, and "demogs" and "items" contain expression-based conditions that those members should satisfy:

- $\forall u \in g.members, \forall \langle a, v \rangle \in g.demogs, \langle a, v \rangle \in u$;
- $\forall u \in g.members, \forall i \in g.items, \exists \langle u, i, t \rangle \in \mathcal{E}$.

For instance, $g_1 = \langle \{u_1, u_2, u_3\}, \{\langle gender, female \rangle\}, \{Titanic\} \rangle$ represents a user group whose members are all females and watched the movie Titanic. We denote the set of all user groups as $\mathcal{G}$.

3 DESIGN CONSIDERATIONS

To achieve a fully-connected pipeline for user group analytics, we need to resolve fragmentations between components, so that analysts can obtain end-to-end group-centric insights from raw user data. Akin to any other data pipeline, UserDEV consists of a "logical layer", which describes components and their inputs/outputs, and a "physical layer", where implementations are hosted. In the logical layer, UserDEV adapts a star-like architecture where all communications are orchestrated by a *central node* (i.e., a visual analytics interface) using a *common data exchange model*. Figure 1 illustrates

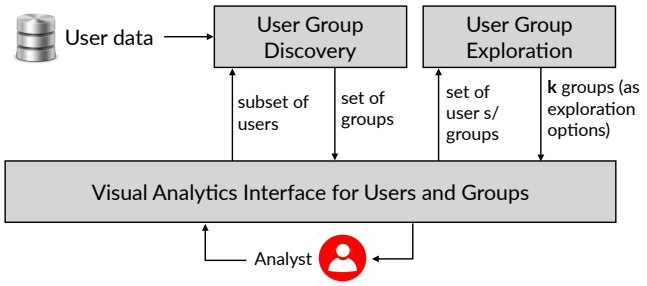

Figure 1: UserDEV architecture.

UserDEV architecture where the visual analytics interface is considered as the central node. In the following, we explain how *common data exchange model* and *visual-centric analytics* ensure that group analytics pipelines are fully connected with reduced switching, learning, and transfer costs.

3.1 Common Data Exchange Model

To reduce transfer cost in pipelines (i.e., **C3**) and ensure genericness (i.e., **C5**), it is necessary to define a common data exchange model to represent the inputs/outputs of all components. Formalizing such a model for group analytics is challenging, because of the different nature of the components. Group discovery has an optimization nature which admits user data as input and generates a group space as output. Group exploration has a focus on selecting a subset of groups which interests analysts. On the other hand, group visualization admits as input a set of groups, and returns their association to visual variables.

We consider the least common atomic concept among all components as the data exchange model, i.e., a *subset of users*, $\mathcal{S} \subseteq \mathcal{U}$. All conversations between the components are reduced to the set $\mathcal{S}$. Discovery communicates its output to exploration and visualization in form of a group set. Also, exploration reports explored options as a group set. Visualization returns a set of users for another round of discovery or exploration. Note that all concepts of "user sets", "groups", "group sets", and even users in their entirety $\mathcal{U}$, are instances of $\mathcal{S}$. In Figure 1, labels on arrows illustrate these concepts. To minimize the transfer cost, we also add some meta data when transferring a set of users from one component to the other, i.e., the distribution of demographics for users inside the subset. This simple yet powerful communication means does not depend on any particular implementation of components in the physical layer, and guarantees genericity of UserDEV. Hence, the components are considered as generic black-boxes whose inputs and outputs are all instances of $\mathcal{S}$.

3.2 Visual-centric Analytics

In most data analytics scenarios, analysts make sense of results using visualization [35]. In UserDEV, we consider the group visualization component as the central node which connects other components together. This is the one destination where both analysts and user group analytics components meet. UserDEV is the first system which combines the power of a visual interface with discovery and exploration of user groups.

In UserDEV, the visualization layer communicates with the analyst to either output a visualized subset of users as the final result, or get as input a subset of users. The input subset can feed either discovery or exploration components. In the former, discovery generates all interesting groups within the input subset. In the latter, exploration finds groups that are relevant to the input subset. In both cases, the output will be delivered to the visualization layer again, in form of a group set (i.e., an instance of $\mathcal{S}$). The set will then be visualized using visual variables. The system will iterate again based on new feedback on the visualized results.

4 FRAMEWORK

UserDEV provides a visualization-centric analytics interface which communicates with group discovery and exploration using a common data exchange model $\mathcal{S}$. In this section, we review the functionality of UserDEV components. We also discuss how UserDEV is used as a visual enabler for user group analytics.

4.1 Group Discovery

User group discovery refers to a set of approaches which derive value from user data in the form of user groups. Discovery is defined as a function $discover(\mathcal{D}, \rho) \rightarrow \mathcal{G}$ which admits as input the user data $\mathcal{D}$ and an optimization objective ρ, and returns the set of groups where $\rho(\mathcal{G}, \mathcal{D})$ is optimized. In [36, 37], several optimization objectives for group discovery are listed, such as frequency, density, coverage, and surprisingness.

To discover user groups in $\mathcal{D}$, different aspects of the data can be used, i.e., demographic attributes $\mathcal{A}$, items $\mathcal{I}$, and similarities $\mathcal{X}$. Attribute-based discovery methods consider users as individual entities and leverage their common demographic attributes to form groups (e.g., discovering the group of young female students) [5, 38, 39]. Item-based discovery mines groups based on common items between users (e.g., discovering a group of users who watched the movie Titanic) [9]. Similarity-based discovery methods leverage connections between pairs of users (e.g., affinity, following, friendship) and divide users into communities with stronger internal connectivity than external connectivity [40–45]

In UserDEV, the analyst defines the subset of users, the optimization function, and the aspect of data that should be used for discovery. The system then returns the set of groups with optimized values on the optimization objective.

4.2 Group Exploration

Exploration refers to a set of approaches which enable interaction with user groups. It is subsumed by HILDA (Human-in-the-Loop Data Analytics) which reflects the involvement of analysts in the data analytics pipeline. Group exploration helps analysts navigate the space of user groups to obtain insights and validate their hypotheses on user data [46]. In many exploration scenarios, the analyst only has a *partial understanding of her needs* and seeks to refine them as she extracts more insights from the data. Exploration is defined as a function $explore(g) \rightarrow \mathbb{P}(\mathcal{G})$ which admits as input a group $g \in \mathcal{G}$ and returns a set of relevant groups to g. Note that $\mathbb{P}(\mathcal{G})$ is the power-set of $\mathcal{G}$.

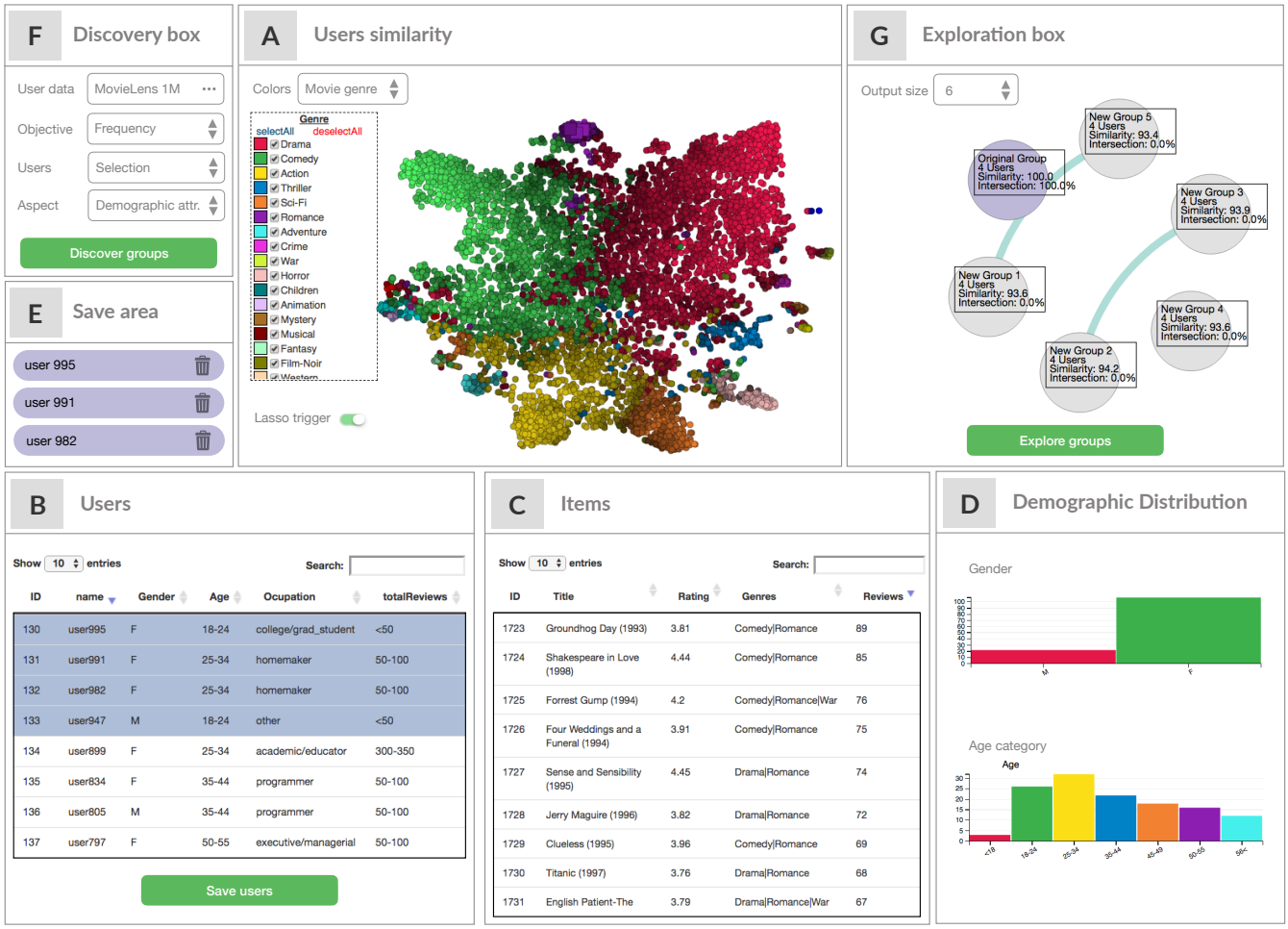

Figure 2: UserDEV visual interface displaying the MovieLens dataset.

At each exploration iteration, the analyst increases her partial understanding of the analysis task. UserDEV captures this awareness in form of "group example", and returns few other groups similar to the provided example, by aggregating similarities in $\mathcal{X}$ [6].

The exploration parameter "output size" (denoted as k) defines how many groups should be returned as exploration options. UserDEV returns by default top-k groups with minimal "overlap" in-between (by minimizing Jaccard similarity between group pairs), to enable analysts investigate different directions. For simplicity, we assume that UserDEV hard-wires an overlap-based exploration [6]. However in practice, other exploration mechanisms can also be employed, e.g., contrast-based exploration [12] and multi-objective exploration [5]. Note that the output of exploration is always an instance of $\mathcal{S}$.

4.3 Group Visualization

Visualization is the central node in UserDEV as it orchestrates the components of user group analytics. At the core of visualizing user groups sits a mapping function $visualize(G \subseteq \mathcal{G}) \rightarrow \mathcal{V}(G)$ which associates G's characteristics (size, overlapping users, items in common, user similarities) to visual variables, i.e., $\mathcal{V}(G)$.

The literature contains very few approaches for visualizing user groups. Traditionally, this is performed directly on raw user data using off-the-shelf visualization products and libraries such as D3, Tableau[3], Spotfire[4], QlikView[5], Gephi[6], and OpenGL[7]. Applied directly on raw data, these solutions are mostly static and do not fully support sophisticated views and exploratory analysis on user groups. More advanced techniques employ graph visualization, where nodes are either users or groups, and edges are weighted using similarity values or overlapping users, respectively [47–49].

Figure 2 illustrates the visual interface of UserDEV applied to MovieLens. The interface has distinct views to display information and statistics of users and groups, and enable access to discovery and exploration. UserDEV uses a coordinated user interface that updates the information displayed after any interaction. Hereafter, we explain different views of this interface, and discuss how the entirety of views addresses concerns of fragmented pipelines.

[3] *Tableau software suite: https://www.tableau.com/*
[4] *Tibco Spotfire: https://spotfire.tibco.com*
[5] *QlikView: https://www.qlik.com/us/products/qlikview*
[6] *Gephi: The Open Graph Viz Platform: https://gephi.org*
[7] *https://www.opengl.org*

A **"Users similarity" view** displays a 2D projection of the similarity values in $\mathcal{X}$ using t-SNE projection [50], which dictates the position of users in the view. This provides an overall view of the user data. UserDEV computes similarities based on commonalities in $\mathcal{E}$. Two users u and u' are maximally similar iff $\forall\langle u,i,t\rangle \in \mathcal{E}, \exists\langle u',i,t\rangle \in \mathcal{E}$. Analysts can color-code users based on an attribute. For instance in Figure 2, users are color-coded based on their dominant genre, i.e., the genre for which they have reviewed movies the most. Analysts have access to a lasso tool which lets them pick a subset of users.

B **"Users" view** lists user's detailed information in form a table whose schema contains demographic attributes $\mathcal{A}$. In case of MovieLens, demographic attributes are gender, age, occupation, and total number of reviews. Users can be sorted according to any column in the table, by clicking on the column. Analysts can also perform search on users using any demographic attribute. By default, the view lists all users. When the lasso tool is used in view **A**, the list will be limited to the users retained by the lasso. The analyst can select one or multiple users in this table and keep them for future investigation (view **E**), by clicking on the "Save users" button.

C **"Items" view** lists items in form of a table whose schema is item's attributes. The view only shows items associated to users in view **B**. Hence a lasso in view **A** restricts the list of items as well. Items can be sorted according to any column in the table, by clicking on the column name. Analysts can also perform search on items using any item attribute.

D **"Demographics distribution" view** displays statistics over the demographic attributes of users. The statistics is only computed for users in view **B**. Hence a lasso in view **A** changes the statistics as well. Analysts can interact with the charts in this view by clicking an attribute value, e.g., the "female" bar in the gender bar chart. This will filter users in all other views to ones which satisfy the selected attribute value. A second click on a selected attribute value will undo the operation.

E **"Save area" view** displays the context of a buffer that lists the users saved in view **B**. Analysts can remove any user from this view or even truncate the buffer to restart the selection process.

F **"Discovery box" view** lets analysts discover groups, by specifying user data, and configuring discovery parameters, i.e., optimization objective, and discovery aspect (demographic attribute, item, similarity, or all of them). User data should be adapted to the model discussed in Section 2. The "Users" parameter lets analysts limit the scope of discovery: the discovery can be done on the entire user data, or the selection shown in view **E**. It is shown in Figure 2 that the analyst requests an attribute-based discovery on a selection of users in MovieLens by optimizing frequency (a common optimization objective in data mining [51]). To avoid information overload, UserDEV does not show all discovered groups $\mathcal{G}$. Instead, it shows a subset only as a result of exploration (i.e. view **G**).

G **"Exploration box" view** is an area for displaying the results of group exploration. Group exploration functions on the set $\mathcal{G}$. Hence before any exploration, at least one round of discovery (view **F**) is required. By default, the input group $g \in \mathcal{G}$ for exploration is the one whose members are shown in view **E**. The analyst can also click on one of the exploration groups to set it as the input group for the next exploration round. The exploration is fired when the button "Explore groups" is clicked.

Views **A** to **G** constitute the visualization component of UserDEV. A valid question is *"how do these views address concerns of fragmented pipelines, collectively?"* The views in UserDEV are dependent on each other and consume common resources. Hence they are aware of each other's situation (addressing **C1**). For instance the buffer in view **E** makes a dependency connection between visualization of users in view **A** and both discovery and exploration, in views **F** and **G**, respectively. On the other hand, the views enable the analysis of users and groups at the same time, hence reducing the switching cost (i.e., **C2**). Also all views communicate only using the common data exchange model $\mathcal{S}$, hence reducing the transfer cost (i.e., **C3**). The coordinated interface of UserDEV facilitates the explainability of results, hence reducing the learning cost. The consequence of each interaction will be immediately displayed on all other views which helps analysts keeps a bird's-eye view on their data (i.e., **C4**). Also thanks to its simple communication model, UserDEV views are independent from any physical implementation of discovery/exploration algorithms, hence a variety of algorithms can be employed (addressing **C5**).

5 USERDEV IN PRACTICE

In this section, we extend the example in Section 1 as a use case and demonstrate how UserDEV helps analysts exploit user groups, in practice. Recall that Emma's goal in our example is to gather a diverse set of reviewers at a first screening of Drama and Comedy movies. We set out to identify a group that contains reviewers with Comedy as their dominant genre, reviewers with Drama as their dominant genre, and a mix of additional reviewers who differ from others either in demographics or in interests. Our use case is defined on the MovieLens 1M dataset where $|\mathcal{U}| = 6,040$, $|\mathcal{I}| = 3,900$ (items are movies), and $|\mathcal{E}| = 1,000,209$ (events are movie ratings).

Group Discovery. Initially, Emma wants to select a subset of users and discover groups in the subset. The traditional way of performing this is to first use Excel or a DBMS to select users, export the subset, re-format it to match the discovery's input, and then run discovery. An alternative way is to use a fully connected pipeline to reduce the burden. In UserDEV, Emma selects relevant users to her search using the lasso tool in view **A**. She then selects users with either Comedy or Drama dominant genres (green-shaded and red-shaded points, respectively) and saves them. The coordinated interface gets updated immediately to only display the information associated with the current selection. Then she requests to discover frequency-optimized item-based groups (in view **F**) using the selection in view **E**. She is now ready to explore discovered groups.

Group Exploration and Visualization. After discovery, Emma needs to explore groups and form her reviewer list. Typically it requires to import discovery results in an off-the-shelf visualization tool and follow several sub-optimal back-and-forth loops. In UserDEV, groups are ready for a visual exploration. Emma looks at views **B** and **C** to find an overall understanding of the subset of users and items she is currently looking at. Using view **D**, she performs selections on the age to only keep 25-34 years old, resulting in 78 reviewers (44 male and 34 female). She also limits the occupation to "academic/educator" resulting in 10 reviewers

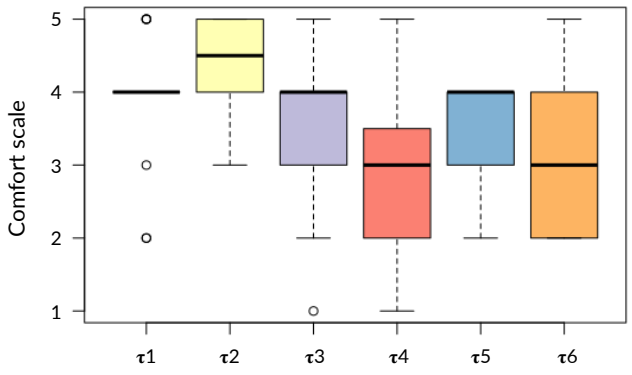

Figure 3: Comfort scales of tasks τ_1 to τ_6.

(6 male, 5 female). These 10 users constitute her input group for exploration. Emma then requests an exploration with $k = 6$ (following output size recommendations in [52]) in view **G**. She will then examine the exploration options to pick interesting reviewers. For instance, she observes that members of 3 out of 6 groups review many Romance-genre movies alongside Comedy and Drama. Emma hand-picks 5 reviewers among those groups whose age > 34 and whose occupation differs from academic/educator (to corroborate diversity), and saves them. Users in the other groups review many Horror, Thriller, and Sci-Fi movies. She hand-picks 5 reviewers from those groups as well, with demographics of age being within 25-34 and occupation being academic/educator. By saving those 5 reviewers, view **E** contains 20 reviewers in total, which serve as Emma's reviewer list.

6 DISCUSSION ON EVALUATION

We described UserDEV, a visualization-centric pipeline for user group discovery and exploration. The principled question is *"how such a group analytics pipeline can be evaluated?"* In this section, we first present a preliminary user study to evaluate the effectiveness of UserDEV. Then we discuss future directions of more extensive evaluations for user group analytics pipelines.

We performed a preliminary user study on MovieLens with 24 participants. Most participants were male (70.83%) with an average age of 30 years old. Also, 51.7% of participants were experts in information visualization, and others were novice. We used the following physical implementations of group analytics components: LCM frequent item-set mining [53] as attribute-based and item-based group discovery (each frequent item-set is a group), IUGA [6] as group exploration, and a web-based GUI implemented in Angular framework as group visualization. Based on the categorization of visualization tasks in [54, 55], we consider six following tasks in increasing order of difficulty: τ_1: "inspect demographics of users who mostly watch romantic movies, and discover groups accordingly", τ_2: "filter users in age and save the smaller set", τ_3: "explore groups with saved users as the input group", τ_4: "identify two highly-reviewed movies which are watched both by the input group and users in explored groups", τ_5: "detect differences in reviews among explored groups", τ_6: "name few genres of users which are different from the ones for the input groups". *Note that the first three tasks focus on the common data exchange model, and the rest on the visual-centric analytics.* We ask participants about their level of comfort (from 1 to 5) in performing these tasks where the scale 5 denotes the easiest.

Figure 3 illustrates the results in form of a box-plot, where tasks τ_1 to τ_6 are on the X-axis and their comfort score is on the Y-axis. We observe that participants are successful in completing tasks τ_1 and τ_2 which relate to filtering user data and discovering groups. An average comfort scale of 4 for τ_1 shows the tight connection between group discovery and visualization. Tasks τ_3 and τ_4 seem to be more challenging as participants require to make group-level interactions while scanning users and movies. An average comfort scale of 3.5 for τ_3 shows the tight connection between group exploration and visualization. The task τ_5 requires visual comparisons among user groups, and our participants succeed to handle that. Concerning the task τ_6, while most participants were able to complete it, many others complained about its difficulty as it needs the verification of several views. In general, an overall comfort scale of 3.62 shows that participants are at ease in performing user group analytics task with UserDEV.

Our user study is preliminary and requires further investigations on the architecture and common data exchange model. Hence the need for a principled usability evaluation arises, which we consider as a future perspective. Despite the established body of related work for evaluating group discovery, group exploration, and group visualization separately, there is no evaluation methodology for their combination [56, 57]. A valid question is whether *we can evaluate UserDEV with a combination of methods proposed to evaluate its components?* We briefly mention two opportunities of all-together evaluation of group discovery, group exploration, and group visualization, i.e., isolation and benchmarking.

Isolation. The most popular approach is to isolate human-oriented aspects of UserDEV (i.e., exploration and visualization) and evaluate the remaining aspects using traditional discovery-based measures, such as execution time and memory usage. For human-oriented aspects, a user study must be designed using crowdsourcing platforms such as Amazon Mechanical Turk[8] and CrowdFlower[9] [58].

Benchmarking. The quality of UserDEV can be assessed by comparisons against standard tests, i.e., benchmarks. Benchmarks are a common practice in the database community (e.g., LDBC social network benchmark [59] and REACT data exploration benchmark [60]). The visualization community still lacks a benchmark and relies on user studies as the only means of evaluation [61, 62]. In [63, 64], a vision towards an interactive visual benchmark is proposed. One of our ongoing directions is to build a benchmark which covers the whole pipeline of user group analytics.

In summary, the evaluation of a user group analytics framework need to go far beyond typical user studies and quantitative measures. Domain-specific benchmarks should be designed to capture human factors in group analytics in an objective fashions and accompany subjective user studies.

ACKNOWLEDGMENT

This work is supported by CDP LIFE project under grant C7H-ID16-PR4-LIFELIG.

[8] *https://www.mturk.com*
[9] *https://www.crowdflower.com*

REFERENCES

[1] Forget the quantified-self, we need to build the quantified-us. http://www.wired.com, 2014.

[2] Connor C Gramazio, Karen B Schloss, and David H Laidlaw. The relation between visualization size, grouping, and user performance. *IEEE transactions on visualization and computer graphics*, 20(12):1953–1962, 2014.

[3] James Doodson, Jeff Gavin, and Richard Joiner. Getting acquainted with groups and individuals: Information seeking, social uncertainty and social network sites. In *ICWSM*, 2013.

[4] Behrooz Omidvar-Tehrani and Sihem Amer-Yahia. Tutorial on data pipelines for user group analytics. In *SIGMOD*, 2019.

[5] Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, Pierre-François Dutot, and Denis Trystram. Multi-objective group discovery on the social web. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, pages 296–312, 2016.

[6] Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, and Alexandre Termier. Interactive user group analysis. In *CIKM*, pages 403–412. ACM, 2015.

[7] Sihem Amer-Yahia, Behrooz Omidvar-Tehrani, João Comba, Viviane Moreira, and Fabian Colque Zegarra. Exploration of user groups in VEXUS. In *34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018*, pages 1557–1560, 2018.

[8] Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. The persuasive power of data visualization. *IEEE transactions on visualization and computer graphics*, 20(12):2211–2220, 2014.

[9] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*, volume 27. ACM, 1998.

[10] Mark EJ Newman. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

[11] Cristian Felix, Anshul Vikram Pandey, and Enrico Bertini. Texttile: an interactive visualization tool for seamless exploratory analysis of structured data and unstructured text. *IEEE transactions on visualization and computer graphics*, 23(1):161–170, 2017.

[12] Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, and Laks V. S. Lakshmanan. Cohort representation and exploration. In *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, pages 169–178, 2018.

[13] Dawei Jiang, Qingchao Cai, Gang Chen, H. V. Jagadish, Beng Chin Ooi, Kian-Lee Tan, and Anthony K. H. Tung. Cohort query processing. *PVLDB*, 10(1):1–12, 2016.

[14] Habibur Rahman, Senjuti Basu Roy, Saravanan Thirumuruganathan, Sihem Amer-Yahia, and Gautam Das. Optimized group formation for solving collaborative tasks. *The International Journal on Very Large Data Bases (VDLBJ)*, 28(1):1–23, 2019.

[15] Safak Aksoy, Eda Atilgan, and Serkan Akinci. Airline services marketing by domestic and foreign firms: differences from the customersâĂŹ viewpoint. *Journal of Air Transport Management*, 9(6):343–351, 2003.

[16] Xiaohua Hu. A data mining approach for retailing bank customer attrition analysis. *Applied Intelligence*, 22(1):47–60, 2005.

[17] Vinicius Machado, Roger A. Leite, Felipe A. Moura, Sergio Augusto Cunha, Filip Sadlo, and João Luiz Dihl Comba. Visual soccer match analysis using spatiotemporal positions of players. *Computers & Graphics*, 68:84–95, 2017.

[18] Olivier Palombi, Fabrice Jouanot, Nafissetou Nziengam, Behrooz Omidvar-Tehrani, Marie-Christine Rousset, and Adam Sanchez. Ontosides: Ontology-based student progress monitoring on the national evaluation system of french medical schools. *Artificial Intelligence in Medicine*, 2019.

[19] Monika Rani, Kumar Vaibhav Srivastava, and Om Prakash Vyas. An ontological learning management system. *Computer Applications in Engineering Education*, 24(5):706–722, 2016.

[20] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248. ACM, 2018.

[21] Kevin Zeng Hu, Diana Orghian, and César A. Hidalgo. DIVE: A mixed-initiative system supporting integrated data exploration workflows. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2018, Houston, TX, USA, June 10, 2018*, pages 5:1–5:7, 2018.

[22] Yixiang Fang, Reynold Cheng, Siqiang Luo, Jiafeng Hu, and Kai Huang. C-explorer: Browsing communities in large graphs. In *Proc. VLDB Endowment*, volume 10, 2017.

[23] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.

[24] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[25] Tilmann Rabl, Jonas Traub, Asterios Katsifodimos, and Volker Markl. Apache flink in current research. *it - Information Technology*, 58(4):157–165, 2016.

[26] Kevin Hu, Diana Orghian, and César Hidalgo. Dive: A mixed-initiative system supporting integrated data exploration workflows. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, page 5. ACM, 2018.

[27] Daniel B Perry, Bill Howe, Alicia MF Key, and Cecilia Aragon. Vizdeck: Streamlining exploratory visual analytics of scientific data. *iSchools*, 2013.

[28] Stef van den Elzen and Jarke J van Wijk. Small multiples, large singles: A new approach for visual data exploration. In *Computer Graphics Forum*, volume 32, pages 191–200. Wiley Online Library, 2013.

[29] Mehmet Adil Yalçın, Niklas Elmqvist, and Benjamin B Bederson. Keshif: Rapid and expressive tabular data exploration for novices. *IEEE transactions on visualization and computer graphics*, 24(8):2339–2352, 2018.

[30] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2016.

[31] Kai Huang, Sourav S Bhowmick, Shuigeng Zhou, and Byron Choi. Picasso: exploratory search of connected subgraph substructures in graph databases. *Proceedings of the VLDB Endowment*, 10, 2017.

[32] Uwe Jugel, Zbigniew Jerzak, Gregor Hackenbroich, and Volker Markl. Faster visual analytics through pixel-perfect aggregation. *Proceedings of the VLDB Endowment*, 7(13):1705–1708, 2014.

[33] Yongjoo Park, Michael Cafarella, and Barzan Mozafari. Visualization-aware sampling for very large databases. In *ICDE*. IEEE, 2016.

[34] Eugene Wu, Leilani Battle, and Samuel R Madden. The case for data visualization management systems: vision paper. *Proceedings of the VLDB Endowment*, 7(10):903–906, 2014.

[35] Jeffrey Heer and Joseph M Hellerstein. Tutorial on data visualization and social data analysis. *Proceedings of the VLDB Endowment*, 2(2):1656–1657, 2009.

[36] Liqiang Geng and Howard J Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9, 2006.

[37] Martin Kirchgessner, Vincent Leroy, Sihem Amer-Yahia, and Shashwat Mishra. Testing interestingness measures in practice: A large-scale analysis of buying patterns. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 547–556. IEEE, 2016.

[38] Mahashweta Das, Sihem Amer-Yahia, Gautam Das, and Cong Yu. Mri: Meaningful interpretations of collaborative ratings. *Proceedings of the VLDB Endowment*, 4(11):1063–1074, 2011.

[39] Jinfei Liu, Li Xiong, Jian Pei, Jun Luo, and Haoyu Zhang. Finding pareto optimal groups: group-based skyline. *Proceedings of the VLDB Endowment*, 8(13):2086–2097, 2015.

[40] Michel Plantié and Michel Crampes. Survey on social community detection. In *Social media retrieval*, pages 65–85. Springer, 2013.

[41] Jungeun Kim and Jae-Gil Lee. Community detection in multi-layer graphs: A survey. *ACM SIGMOD Record*, 44(3):37–48, 2015.

[42] Giulio Rossetti and Rémy Cazabet. Community discovery in dynamic networks: a survey. *ACM Computing Surveys (CSUR)*, 51(2):35, 2018.

[43] Steve Harenberg, Gonzalo Bello, L Gjeltema, Stephen Ranshous, Jitendra Harlalka, Ramona Seay, Kanchana Padmanabhan, and Nagiza Samatova. Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):426–439, 2014.

[44] Jure Leskovec, Kevin J Lang, and Michael Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web*, pages 631–640. ACM, 2010.

[45] Michelle L Gregory, Dave W Engel, Eric Belanga Bell, Andy Piatt, Scott Dowson, and Andrew J Cowell. Automatically identifying groups based on content and collective behavioral patterns of group members. In *ICWSM*, 2011.

[46] Robert West and Jure Leskovec. Automatic versus human navigation in information networks. In *ICWSM*, 2012.

[47] Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[48] Ivan Herman, Guy Melançon, and M Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on visualization and computer graphics*, 6(1):24–43, 2000.

[49] Danai Koutra, Di Jin, Yuanchi Ning, and Christos Faloutsos. Perseus: an interactive large-scale graph mining and visualization tool. *Proceedings of the VLDB Endowment*, 8(12):1924–1927, 2015.

[50] L.J.P van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9: 2579âĂŞ2605, Nov 2008.

[51] Ramakrishnan Srikant and Rakesh Agrawal. Mining generalized association rules. *ACM*, 1995.

[52] George Miller. Human memory and the storage of information. *IRE Transactions on Information Theory*, 2(3):129–137, 1956.

[53] Takeaki Uno, Masashi Kiyomi, and Hiroki Arimura. Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *Fimi*, volume 126, 2004.

[54] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[55] David Gotz and Michelle X Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009.

[56] Behrooz Omidvar-Tehrani and Sihem Amer-Yahia. Tutorial on 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*, pages 2307–2308, 2018.

[57] Lilong Jiang, Protiva Rahman, and Arnab Nandi. 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*, pages 1637–1644, 2018.

[58] Anand Inasu Chittilappilly, Lei Chen, and Sihem Amer-Yahia. A survey of general-purpose crowdsourcing techniques. *IEEE Transactions on Knowledge and Data Engineering*, 28, 2016.

[59] Orri Erling, Alex Averbuch, Josep Larriba-Pey, Hassan Chafi, Andrey Gubichev, Arnau Prat, Minh-Duc Pham, and Peter Boncz. The ldbc social network benchmark: Interactive workload. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 619–630. ACM, 2015.

[60] Tova Milo and Amit Somech. Next-step suggestions for modern interactive data analysis platforms. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 576–585. ACM, 2018.

[61] Tamara Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928, 2009.

[62] James J Thomas. *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005.

[63] Philipp Eichmann, Emanuel Zgraggen, Zheguang Zhao, Carsten Binnig, and Tim Kraska. Towards a benchmark for interactive data exploration. *IEEE Data Eng. Bull.*, 39(4):50–61, 2016.

[64] Leilani Battle, Marco Angelini, Carsten Binnig, Tiziana Catarci, Philipp Eichmann, Jean-Daniel Fekete, Giuseppe Santucci, Michael Sedlmair, and Wesley Willett. Evaluating visual data analysis systems: A discussion report. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2018, Houston, TX, USA, June 10, 2018*, pages 4:1–4:6, 2018.