Toward Interactive User Data Analytics

Sihem Amer-Yahia

CNRS, University Grenoble-Alps, France

Abstract. User data can be acquired from various domains. This data is characterized by a combination of demographics such as age and occupation and user actions such as rating a movie, reviewing a restaurant or buying groceries. User data is appealing to analysts in their role as data scientists who seek to conduct large-scale population studies, and gain insights on various population segments. It is also appealing to novice users in their role as information consumers who use the social Web for routine tasks such as finding a book club or choosing a restaurant.

User data exploration has been formulated as identifying group-level behavior such as *Asian women who publish regularly in databases*. Group-level exploration enables new findings and addresses issues raised by the peculiarities of user data such as noise and sparsity. I will review our work on one-shot [1–4] and interactive [5] user data exploration. I will then describe the challenges of developing a visual analytics tool for finding and connecting users and groups.

1 Data Model

Given a set of users U and a set of items I, we define *user data* as a database D of tuples $\langle u, i, val \rangle$ which represent a value *mathitval* induced by an action such as browsing, tagging, rating, and tweeting, of user u, from a set U, on item i, from a set I. For instance, the tuple $\langle John, Titanic, 4 \rangle$ describes that John rated the movie Titanic with a score of 4. The tuple $\langle Tiffany, tweet_{id}, Hemophilia \rangle$ describes that Tiffany tweets about 'Hemophilia'.

Users and items have attributes drawn from a set A. The set of user attributes (age, gender, diet, occupation, etc.) is denoted as $A_u \subset A$ and the set of item attributes (book author, movie director, tweet language, etc.) is denoted as A_i where $A_u \cup A_i = A$ and $A_i \cap A_u = \emptyset$. Each attribute $a_i \in A$ has a set of values $dom(a_i) = \{v_1^i, v_2^i \ldots\}$. V denotes the set of all attribute values.

Multiple datasets could be represented in our model. We have been using over 5B tweets, about 300M customer receipts from a retail chain of 1,800 stores, 10M rating records from MovieLens, 50M artist ratings from LastFM, and about 200K book ratings from BookCrossing.

User Group. A user group g, is a subset of U to which is associated a description defined as $[v_1^1, v_2^1 \ldots v_j^i \ldots]$ where each $v_j^i \in dom(a_i)$ either holds for all users in g (if $a_i \in A_u$) or holds for all items in g (if $a_i \in A_i$). For instance,

the group $[25, student, action]$ contains 25-year old students who watch action movies. Here, "25" $\in dom(age)$, "$student$" $\in dom(occupation)$ and "$action$" $\in dom(genre)$ where $\{age, occupation\} \subset A_u$ and $genre \in A_i$. G refers to the set of all user groups. We also define two functions $users(g)$ and $desc(g)$ that return g's members and g's description, respectively. $|G| = 2^{|V|+|I|}$, which can be very large even with a few attribute values and items. For example, with $|I| = 5$, $|A| = 4$ and 5 values per attribute (i.e., $|V| = 20$), $|G|$ will be in order of 10^7.

2 One-Shot Exploration

We describe motivating examples and then summarize our contributions and takeaways for one-shot user data exploration.

Example 1 (Finding a Suitable Movie). Sofia wants to see a comedy. IMDb gives her is a global average rating or an average rating broken down by pre-packaged user segments. Sofia would like to select a set of comedies and only wants to see segments where users have rated those comedies similarly. Those segments, akin to user groups, must be mined from the rating records of her selected comedies and cannot be pre-packaged for every possible subset of D.

Example 2 (Quantified-Self). Mary[1] is an avid book reader and is very active on BookCrossing. She has over 1,000 ratings (ranging from 1 to 10 but mostly high) for her favorite author, *Debbie Macomber*. She is looking to join an online book club where she can find people with whom she agrees and disagrees to engage in a stimulating debate. Returning user groups that highly agree or disagree with Mary would be useful to her.

Summary of Approaches. The first collection of papers state the problem of user data exploration as a one-shot optimization where the input is any subset of the database D and the output is k user groups in G. In [2], we stated the problem as finding k groups whose ratings are uniform or polarized. This could help us solve Sofia's problem in Example 1. Our follow up work [3], examined the case where the k groups are characterized by similar/dissimilar tagging actions. In [1], the input to the problem was stated as finding a set of groups whose rating distribution was close to one distribution provided as input. This could help us solve Mary's problem in Example 2. Finally, in [4], we studied a multi-objective formulation of the problem of finding user groups where criteria such as the coverage of input data, the rating distribution of individual groups, or the diversity of returned groups, could be optimized together.

Summary of Takeaways. In this series of papers, we showed that it is useful to state user data exploration as the problem of optimizing local and global criteria. Individual group size, the distribution of ratings within a group, or the length of a group's description are examples of local criteria. Global criteria refer to the coverage of input records or the divesity of returned groups. We showed the hardness of our problems and proposed heuristics to address them efficiencly.

[1] *We do not know Mary personally but she is a real user on* BookCrossing.

3 Interactive Exploration

We describe a motivating example and then summarize our contributions and takeaways for interactive user data exploration studied in [5].

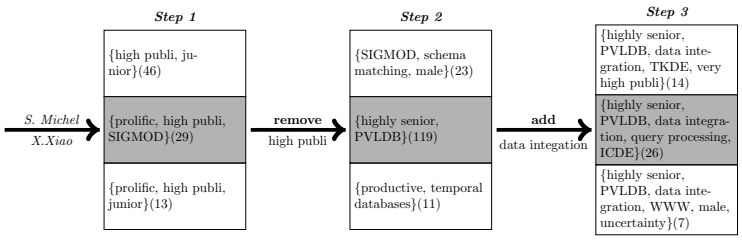

Fig. 1. WebDB 2014 Program Committee Formation

Example 3 (Expert-Set Formation). Martin[2] is looking to build a program committee formed by geographically distributed male and female researchers with different seniority and expertise levels. Martin starts with 2 junior researchers, S. Michel and X. Xiao (Figure 1). The system uses them to discover 3 groups out of which the group described as *prolific*, *high publications* and publishing in *SIGMOD*, has 29 geographically-distributed and gender-distributed researchers. Martin chooses L. Popa, A. Doan, M. Benedikt, and S. Amer-Yahia in those groups. Step 3 reveals 3 other groups out of which 1 group contains 119 highly senior researchers and can be broken into 3 sub-groups whose researchers have expertise in *data integration*. In particular, the group labeled *query processing*, *PVLDB* and *ICDE* contains 26 senior researchers out of which 8 are of interest to the PC chairs: J. Wang, F. Bonchi, K. Chakrabarti, P. Fratenali, D. Barbosa, F. Naumann, Y. Velegrakis and X. Zhou. At this stage and after 3 steps only, Martin covered 80% of the WebDB PC.

Summary of Approach. We developed a framework where, at each iteration, an analyst visualizes k groups, chooses one group of interest, and takes an action on that group (add/remove members, modify description). The analyst then chooses an operation to discover k diverse groups that are relevant to the current group. We provide two opeations: exploit generates k diverse groups that cover the current group, and explore that generates k diverse groups that overlap with the current group. We showed that 50% of the program committees of conferences such as SIGMOD and VLDB and CIKM, can be built in fewer than 9 interactions. We also showed that 80% of the program committee of SIGMOD can be built in 10 steps and that it is closer to 15 for CIKM. The diversity

[2] *Martin Theobald was indeed the* WebDB *PC chair in 2014!*

of topics in CIKM increases the needed steps to cover more committee members.

Summary of Takeaways. In this work, we showed the usefulness of interactive group exploitation and exploration and formulated them as optimization problems. We proved their hardness and devised greedy algorithms to help analysts navigate in the space of groups and reach one or several target users. One important takeaway in interactive exploration is the need for a principled evaluation methodology. That requires the definition of objective measures such as the number of steps necessary to find a single user or a set of users as for Martin in Example 3. It also requires the careful design of appropriate user studies.

4 Visual Analytics

In many exploration scenarios, the analyst only has a partial *partial understanding of her needs* and needs to refine them as she extracts more insights from the data. Ideally, she would be immersed in an environment where she could provide any kind of input related to her needs be it a dataset, rating distributions or a query of interest. The proposed environment would take her input and start suggesting user groups along with some analytics. Such a tool would provide the ability to find and connect user groups in a way that seemlessly integrates our contributions described in Sections 2 and 3. User groups could be visualized in a directed force layout to prevent visual clutter. Histograms and charts that show detailed statistics about groups could be provided. Those statistics must be displayed in coordinated views where a brush on one (e.g., histogram) updates all others instantaneously. In addition to exploit/explore primitives, there is a need to provide long jumps as well as the ability of undoing a previous step. An essential element of this tool would be to let analysts provide explicit feedback in addition to the implicit feedback gathered from their various actions. Feedback will help to customize the exploration by recommending subsequent exploration steps. It will also help evaluate the usefulness of such a tool.

References

1. S. Amer-Yahia, S. Kleisarchaki, N. K. Kolloju, L. V. Lakshmanan, and R. H. Zamar. Exploring rated datasets with rating maps. In *WWW, Perth, Australia*, 2017 (to appear).
2. M. Das, S. Amer-Yahia, G. Das, and C. Yu. MRI: meaningful interpretations of collaborative ratings. *PVLDB*, 4(11):1063–1074, 2011.
3. M. Das, S. Thirumuruganathan, S. Amer-Yahia, G. Das, and C. Yu. An expressive framework and efficient algorithms for the analysis of collaborative tagging. *VLDB J.*, 23(2):201–226, 2014.
4. B. Omidvar-Tehrani, S. Amer-Yahia, P.-F. Dutot, and D. Trystram. Multi-objective group discovery on the social web. In *ECML/PKDD*, pages 296–312. Springer, 2016.
5. B. Omidvar-Tehrani, S. Amer-Yahia, and A. Termier. Interactive user group analysis. In *CIKM*, pages 403–412. ACM, 2015.