ARViz: Interactive Visualization of Association Rules for RDF Data Exploration

Aline Menin, Lucie Cadorel, Andrea G. B. Tettamanzi, Alain Giboin, Fabien Gandon, Marco Winckler

ARViz: Interactive Visualization of Association Rules for RDF Data Exploration

1st Aline Menin
University Côte d'Azur, Inria, CNRS, I3S (UMR 7271)
France
aline.menin@inria.fr

2nd Lucie Cadorel
University Côte d'Azur, Inria, CNRS, I3S (UMR 7271)
France
lucie.cadorel@inria.fr

3rd Andrea Tettamanzi
University Côte d'Azur, Inria, CNRS, I3S (UMR 7271)
France
andrea.tettamanzi@inria.fr

4th Alain Giboin
University Côte d'Azur, Inria, CNRS, I3S (UMR 7271)
France
alain.giboin@inria.fr

5th Fabien Gandon
University Côte d'Azur, Inria, CNRS, I3S (UMR 7271)
France
fabien.gandon@inria.fr

6th Marco Winckler
University Côte d'Azur, Inria, CNRS, I3S (UMR 7271)
France
marco.winckler@inria.fr

***Abstract*—Association rule mining often leads the analyst into a rough rummaging process to identify rules that are relevant to understand specific problems. We propose a visualization interface to assist the rule selection process and evaluate it on an RDF knowledge graph derived from the COVID-19 Open Research Dataset. The user interface supports data exploration with focus on the overview of rules through a scatter plot, subsets of rules through a chord diagram chart, and itemsets through an association graph which is dynamically created by entering an item of interest (i.e. a named entity). Further, the analyst can interactively recover a list of publications containing the named entities involved in a particular rule. Among the original aspects of our approach, we highlight the representation of attributes describing measures of interest (i.e. confidence and interestingness), a visual indication of existence (or not) of symmetry in association rules, the exploration of subsets of rules according to clusters of publications and named entities, and an interactive prompting that aims at expanding the discovery of named entities within selected association rules. We assess our approach through a semi-structured interview involving experts in the domains of data mining and biomedicine, whose feedback could assist the refinement of the visual and interaction tools.**

***Index Terms*—association rule mining, interactive visualization, knowledge graph visualization, RDF visualization, COVID-19**

I. Introduction

The widespread use of Linked Open Data (LOD) for publishing data on the Web has promoted a fast growth of available data sources that provide access to huge amounts of data ever more diversified, which analysis provides valuable information to support decision-making processes in various application areas [1]. In the context of the COVID-19 pandemics, for instance, the whole scientific community got together in a common effort to study, understand, and fight the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), generating a huge amount of scientific publications (over 200,000 and increasing) about coronaviruses and related diseases. Studying and extracting relevant information from this deluge of data is impossible without computational power. Thus, there are multiple efforts towards analyzing and mining the COVID-19 scientific literature by diverse tools and purposes. A promising approach [2] consists on using named-entity (NE) linking to enrich the data by assigning a unique identity to entities (e.g. locations or diseases) mentioned in the text, which are then represented through RDF (Resource Description Framework) knowledge graphs (KGs), a widely used data format to describe the semantics of real-world entities and their relations, and to link the descriptions to further information in semantic LOD repositories [3].

In this paper, we propose an approach to assist the exploration of data from named entities knowledge graphs based on the joint use of association rule mining and visualization techniques. The former is a widely used data mining method to discover interesting correlations, frequent patterns, associations or casual structures among transactions in a variety of contexts. An association rule is an implication of the form $X \rightarrow Y$, where X is an antecedent itemset and Y is a consequent itemset, indicating that transactions containing items in set X tend to contain items in set Y.

Although the approach helps to reduce and focus the exploration of large datasets, analysts are still confronted with the inspection of hundreds of rules in order to grasp valuable knowledge. Moreover, when extracting association rules from named entities knowledge graphs, the items are NEs that form antecedent $\rightarrow$ consequent links, which the user should be able to cross to recover information. In this context, information visualization can help analysts to visually identify interesting rules that are worthy of further investigation, while

providing suitable visual representation to communicate the relationships between itemsets and association rules. Thus, the main contributions of this work are:

- A comparative review of existing visualization interfaces/techniques for association rule exploration regarding task support;
- A visualization interface to assist the exploration of association rules over RDF knowledge graphs and their measures of interest, supporting tasks of comparison, identification, and overview of items and rules;
- A feature to recover/access the data source (e.g. objects of the RDF knowledge graph) through named entities involved in a particular association rule; and
- A formative evaluation of the tools with expert users in Semantic Web and biomedical research.

The remaining of this document is organized as follows. Section II surveys previous works on visualization of association rules. Section III states the problem while explaining the association rules mining process embedded in the visualization pipeline. Section IV presents the design rationale of our approach and the visualization interface. Section V describes the evaluation performed with expert users to assess the usability and suitability of our approach. Section VI discusses the results and Section VII concludes and unveils the limitations and future research perspectives.

II. Related Work and Comparative Review of Techniques

Based on a comprehensive review of the literature, we identified a set of exploratory tasks aimed at assisting analysts of association rules: i) get an overview of rules available, ii) identify interesting rules, iii) identify relationship between items behind rules, iv) identify frequent itemsets, v) compare rules, and vi) recover detailed information that describe the rules (e.g. measures of interest, data source, dataset size). Table I presents the level of support for these exploratory tasks by the visualizations proposed in previous works.

TABLE I
SUMMARY OF RELATED WORK: TASK SUPPORT BY PROPOSED VISUALIZATION INTERFACES/TECHNIQUES. LEGEND: APPROPRIATE, LIMITED SUPPORT, NO SUPPORT.

Ref.	Type of Task					
	Overview	Interesting rules	Items Relations	Frequent itemsets	Compare rules	Details on demand
[4]						
[5]						
[6]						
[7]						
[8]						
[9]						
[10]						
[11]						
[12]						
[13]						

Scatter plots in [4] support the overview of rules and the identification of interesting ones by encoding support and confidence in the x, y axes, and providing zooming, selection and linking operations to enable users to focus on a rule or subset of rules. CrystalClear [11] uses a matrix where rows and columns represent values of confidence and support, and the cells group rules with equal values of those measures. While aggregating rules reduces visual cluttering, gives an overview of the dataset and visually classify rules according to measures of interest, it fails on representing antecedent/consequent items, which prevents the identification of relations between items or itemsets. Also, since every group of rules, regardless of its size, is encoded through identical symbols in terms of color and dimensions, it can mislead the interpretation of rules distribution over measures of interest.

Circular graphs in [5] represent association rules through line segments, which color (blue-to-yellow gradient) encodes direction from antecedent to consequent items placed over the circle's circumference. Although relations between items are easily spotted, the technique does not represent measures of interest hindering the identification of interesting rules.

Parallel coordinates plot in [6] represent association rules through line segments linking items displayed on vertical axes evenly spaced horizontally. Each rule contains two polylines connecting antecedent and consequent items, which thickness and color encode support and confidence, respectively. In [7] they are combined with a directed graph view, which represents the relationship among items or rules through directed links which color and length represent support and confidence, respectively. In both techniques, the amount of line segments increases as the number of association rules augments, which causes visual cluttering and hinder data exploration. Furthermore, one must combine two independent visual coding (color and length) to identify interesting rules, which could engender cognitive overhead. The graph view in [13] represents rules with focus on itemsets, where selection operations allow the user to highlight rules and to explore smaller subsets of rules based on frequent itemsets.

A matrix in [8] provides overview of all rules in the dataset by representing antecedent and consequent items as rows and columns, and rules as the intersection cells, which color (gray scale gradient) encodes the support measure. The matrix supports scaling and scrolling actions to explore more data than the screen space allows, and selection operations to define subsets of items, which corresponding rules can be further explored in a graph view. The latter represents the selected antecedent/consequent items as squares over two vertical axes and rules as circles over a vertical central axis. Squares and circles are connected through line segments, which color and thickness encode rules' support and confidence, respectively.

In [12] rules a represented via a scatter plot, a matrix and a graph. The scatter plot displays the distribution of rules over the intersection of axes encoding support and confidence. The rows and columns of the matrix correspond to antecedent and consequent items, while the crossing cells are colored to represent the lift (a measure of performance) of the rule. Inversely to how rules are often encoded in graph views, i.e. through the edges connecting nodes, the authors propose to represent items and rules as nodes, which color encodes lift values, and

antecedent	consequent	confidence	interestingness	support	isSymmetric	cluster
ritonavir	lopinavir	0.89	0.59	0.0003414	×	label_cluster2
insect	baculovirus	0.71	0.36	0.0024765	×	label_cluster2 article_cluster1
steroid, osteoblast	saon	0.83	0.83	0.0002134	×	no_clustering
duchenne muscular dystrophy	muscular dystrophy	1	1	0.0052854	✓	article_cluster1

Fig. 1. Sample of the dataset of association rules.

the edges connect the nodes to represent their relationship. This approach provides an overview of the rules but hinders the identification of itemsets and rules in large datasets, since no specific order is imposed to the nodes. Furthermore, the authors state that only 100 rules may be visualized via the graph view without cluttering the visualization.

In [9] association rules are represented as vertical line segments embedded with circular dots connecting antecedent and consequent items, which are represented through horizontal lines placed above and below the x-axis. The support of rules is encoded by the order of vertical lines, which color may encode confidence or support. This technique favors the exploration of association rules based on an item or itemsets, while prompting other associated items. However, it cannot visualize more than a few rules simultaneously. InterVisAr [10] allows to explore a set of rule at the time, defined by the consequent item or itemset. The technique is a sort of dynamic bar chart, which bars change according to a selection operation. These allow to investigate how confidence and support are affected by the addition or removal of antecedent items. However, one cannot explore multiple rules simultaneously, preventing comparison and discovery of new rules.

Although there are numerous visualization techniques and interfaces that assist the exploration of association rules, to the extent of our knowledge there is no technique or interface that supports the whole set of tasks. In this paper we provide a visualization interface that combines three visualization techniques to assist the exploration of association rules that support these tasks.

III. Problem Statement

A. Data, Dataset and Mining Process

In this study, we analyze the *Covid-on-the-Web* RDF knowledge graph derived from the 7^{th} version of the CORD-19 corpus [14], which gathers over 50,000 scientific publications about COVID-19, SARS-CoV-2, and related coronaviruses. We focus in one of the several named graphs embedded in the *Covid-on-the-Web*: the *CORD-19 Named Entities Knowledge Graph* (CORD19-NEKG) that contains named entities identified and disambiguated by NCBO BioPortal annotator [15], Wikidata Entity-fishing [16] and DBpedia Spotlight [17]. Using the algorithm proposed in [18], we extract a set of association rules from the CORD19-NEKG, particularly using the NEs identified by Wikidata Entity-fishing tool.

The mining process defines transactions as publications and itemsets as named entities, and uses four clustering approaches (i.e. no clustering, items clustering, transactions clustering, or both items and transactions clustering), which results in four sets of association rules. These undergo a filtering process based on the following criteria: redundancy (i.e. $A, B, C \rightarrow D$ is redundant if $Conf(A, B \rightarrow D) \geq Conf(A, B, C \rightarrow D)$), minimal confidence (i.e. the probability of finding the item Y in a transaction, knowing that the item X is in the same transaction; threshold of 0.7), and interestingness (threshold of 0.3), which thresholds were empirically chosen based on the number of resulting rules.

Particularly, the typical support-based approach (i.e. based on the probability of finding the items X and Y in a transaction) for association rules mining, which fails on discovering interesting rules with low support and over-represents itemsets with high support [18]. Hence, the interestingness measure helps to determine the serendipity of the rule, which value is computed based on the rule's support (Eq. 1) and allows to penalize rules with high incidence of antecedent and/or consequent itemsets in the database. For instance, the NE "virus" appears thousands of times in the dataset and, since the rules are calculated by the number of co-occurrences between two named entities in the publications, it will appear in several rules. However, this information has low relevance for an expert user since the topic of all articles surrounds a type of virus. Thus, the algorithm penalizes rules that have the NE "virus", which does not mean deleting: it is possible that "virus" is associated with a rarer NE, which resulting rule may have an interestigness measure above the established threshold.

$$Inter(X \rightarrow Y) = \left(\frac{Supp(X \rightarrow Y)}{Supp(X)}\right) \times \left(\frac{Supp(X \rightarrow Y)}{Supp(Y)}\right) \times \left(1 - \frac{Supp(X \rightarrow Y)}{\texttt{Total no. of transactions}}\right) \quad (1)$$

The resulting data is arranged in a table (see Fig. 1), where rows correspond to association rules and columns describe antecedent and consequent itemsets, the support, confidence and interestingness measures, whether the rule is symmetric, and the cluster to which the itemsets or transactions belong.

B. Analytical Tasks and Operation Model

The exploratory analyses of association rules are typically focused on (1) *items*, used to find and/or describe the rules involving a particular item, or (2) *rules*, allowing users to (i) explore distinguishable association rules in a dataset, and (ii) identify rules that are worth saving for knowledge acquisition.

Selection	Example of Task
Cluster(s)	Explore rules in a cluster or a set of clusters of items or transactions.
Itemset	Explore existing rules involving the selected items.
Item	Explore the rules involving an item and identify associated items.
Confidence	Explore the rules within a particular range of confidence score.
Interestingness	Explore the rules within a particular range of interestingness score.

Fig. 2. Examples of analysis tasks according to selection operations.

ARViz has three operation types: filtering, sorting, and hovering. Filtering operations allow users to specify subsets of interest from the whole input dataset, which supports a variety of types of tasks such as illustrated in Fig. 2. The user can define subsets by selecting: clusters of items, transactions, or the combination of both, revealing association rules within set of named entities covering a particular research topic of interest; a particular item or itemset, which reveal their associated items; and measures of confidence and interestingness, focusing on stronger or weaker rules.

The sorting operations intend to provide quick visual identification of rules or itemsets of interest by sorting named entities by alphabetic order or according to the number of associated rules, and rules according to measures of confidence or interestingness. Finally, the hovering operations enable the highlighting of rules or itemsets in each visualization technique, easing information retrieval.

IV. Interactive Visualization

For supporting the above-mentioned tasks, we designed ARViz (Fig. 3), a visualization interface for association rules exploration through three synchronized views: a scatter plot chart, a chord diagram and a directed graph. We follow an uniformity principle, providing the same visual encoding for measures of interest and interaction operations across views to support comfortable and coherent user experience.

Our visualization interface is based on a minimalist design, which presents one view at the time to maximize screen space and provides three small lateral buttons to allow the user to display useful panels to manipulate the data only when needed. These panels contain the legends of colors and patterns used for encoding rules (Fig. 3a), the filtering forms allowing the selection of smaller and meaningful subsets (Fig. 3b), and the sorting forms allowing the user to modify the order of how rules and terms are displayed on the chord diagram and the association graph (Fig. 3c). We use a tab-based display allowing the user to switch between views: the overview of rules (Fig. 3d); the circular paginated view of subsets (Fig. 3e); and the exploratory graph view of items (Fig. 3f).

According to the algorithm used in this study, we visually encode measures of confidence and interestingness, instead of support. Since one should interpret both measures together to determine the interest of a rule, we use a bivariate legend that combines two color scales: one per measure (Fig. 3a). The resulting color scale indicates when both measures are high, medium, and low, and when one of them is more important than the other according to the color tendency (i.e. when color leans to rose, it means that the confidence of the rule is higher than its interestingness, and when color leans to green, it indicates otherwise). Finally, the symmetry of rules is represented through a pattern of white circles.

A. Overview of Rules

Although scatter plots present shortcomings, such as visual clutter, they remain powerful to visualize rules distribution over confidence and interestingness. The chart's x-axis represents values of interestingness, while the y-axis represents values of confidence. The crossing points between both measures are represented by diamond symbols, following our uniformity principle, which color encodes measures of interest, texture encodes symmetry, and size encodes the number of rules for each pair $< confidence, interestingness >$ (Fig. 4). Often, the dataset contains sets of rules with the same measure of confidence and interestingness, which only part is symmetric. Thus, we place the diamonds representing symmetric rules on the top of the set of non-symmetric rules. To prevent misinterpretation of information, we provide a tooltip over each diamond displaying the actual number of rules represented, the values of measures of interest, and whether the rules are symmetric or not.

B. Circular Paginated View of Subsets

Circular graphs provide an alternative to complex matrices as the connections between itemsets can be represented by arrows, guiding user's attention. Thus, we propose the circular paginated view of subsets (Fig. 5) to assist the exploration of subsets of rules and the resolution of rule-based tasks. Since the antecedent and consequent sides of association rules can contain more than one item, we adapted the traditional implementation of the diagram to merge multiple ribbons into one or fork one ribbon into multiple ones. The order of arcs around the circumference can be modified by sorting the named entities by alphabetic order or according to the number of association rules involving each item. The latter provides a cleaner view of rules, since the items within the most association rules tend to be related to items that are themselves associated to a large number of rules. Thus, the items will be closely placed reducing the number of ribbons crossing each other.

Displaying a number of rules higher than a certain threshold would clutter the visualization causing overhead when distinguishing colors to determine the confidence, interestingness and symmetry of rules. Therefore, we adopted a "pagination" approach (Fig. 5b) that allows users to browse over subsets of a given number of rules determined and customizable by the user at any time. The same method can be used to browse rules based on subsets of items (e.g. explore the rules involving a subset of 20 items). The technique can display up to 150 rules without cluttering the visualization, i.e. still supporting visual identification of interesting rules.

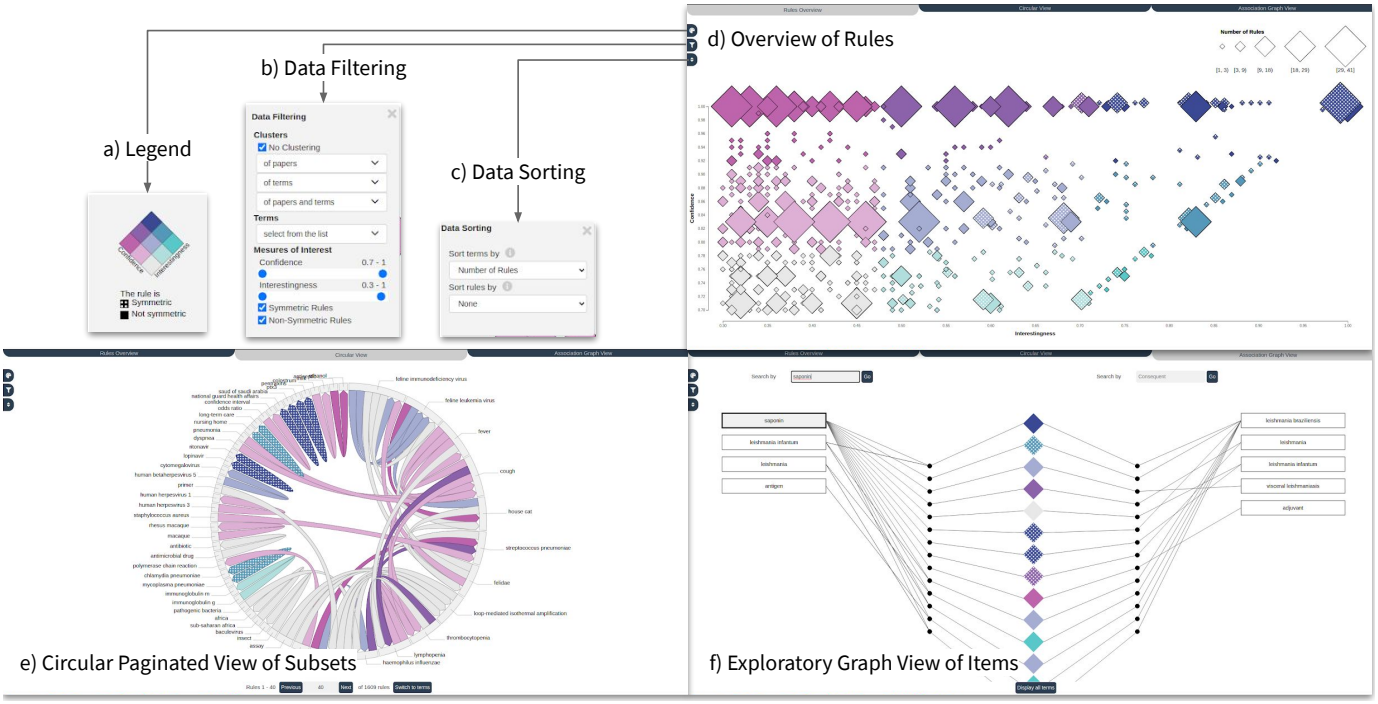

Fig. 3. Interactive visualization: the buttons in the top left corner allow to access panels to view the legend (a), the data filtering (b) and sorting (c). The interface embeds 3 visualizations: a scatter plot to provide overview of rules (d); a chord diagram to provide a paginated exploration of subsets of rules (e); and an association graph to provide an exploratory view of items (f).

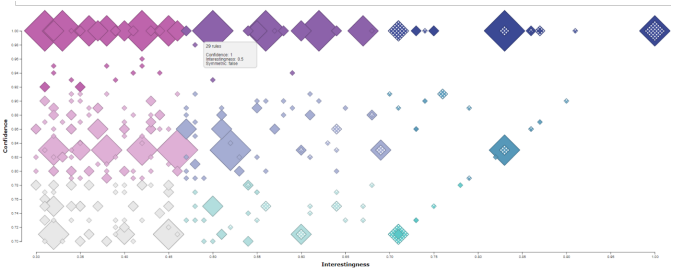

Fig. 4. A scatter plot gives an overview of all the rules while serving as filtering tool to enable the exploration of subsets of rules through the circular paginated view.

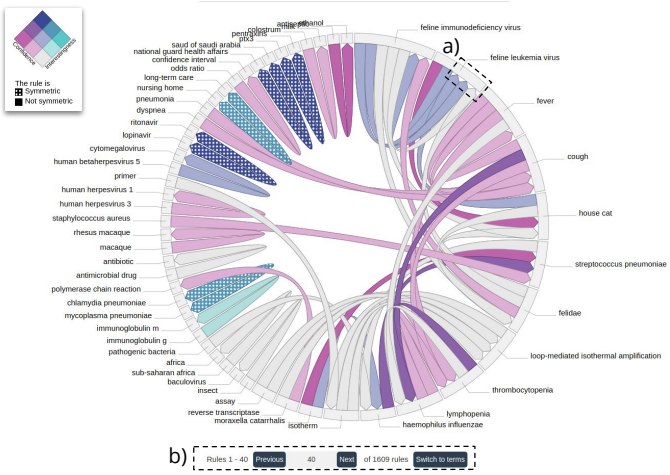

Fig. 5. The circular view: arcs encode single items and ribbons represent the rules connecting them, which color encodes confidence, interestingness, and symmetry. The arrowhead of ribbons (a) encodes the rule direction and points the consequent item. The browsing panel (b) allows to modify the size of subsets and browse between them.

C. The Association Graph View

Association graphs (Fig. 6) give an intuitive portray of antecedent and consequent items involved in rules by representing them as nodes placed in the left and right side of rules. In our design, we represent items over two vertical stacks of labeled rectangles at the left and right extremities of the window, and rules as diamond-shaped nodes placed at the center of the visualization space. The bootstrapping process of the association graph view depends on a trigger action of the user, described hereafter.

Upon selection of an item from the antecedent/consequent lists, the selected item and its related items are placed on the top of the list (Fig. 6b-c), followed by the remaining items available in the dataset (Fig. 6d), colored in gray to indicate their disconnection with the visible rules. **Upon choosing an item via the search bar** (Fig. 6a), the system will display the selected and its related items on the corresponding side and display the associated rules. This process is also triggered **upon the choice of an item via a context menu** embedded on the arcs of the circular view. Continuous interaction is available enabling the user to click on any of the visible items (including the ones in gray), triggering a re-positioning and re-coloring of items. Furthermore, the segments between the

vertical axis of antecedent/consequent and the rules contain a black dot indicating the union of two or more items as an itemset acting as antecedent/consequent (Fig. 6e).

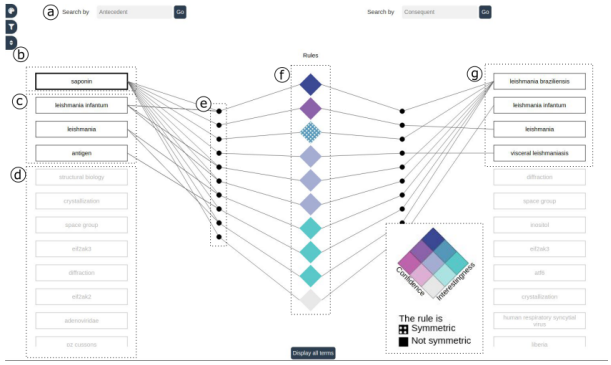

Fig. 6. The association graph view: (a) the search bars to find NEs; (b) the selected antecedent NE; (c) other NEs embedded in the antecedent of rules containing the selected NE; (d) the list of remaining antecedent NEs available in the dataset; (e) junction dots linking items belonging to the same itemset; (f) the color-encoded shapes representing the rules associating both antecedent and consequent NEs; and (g) the consequent NEs of selected rules.

D. Details-on-Demand and Data Source Information

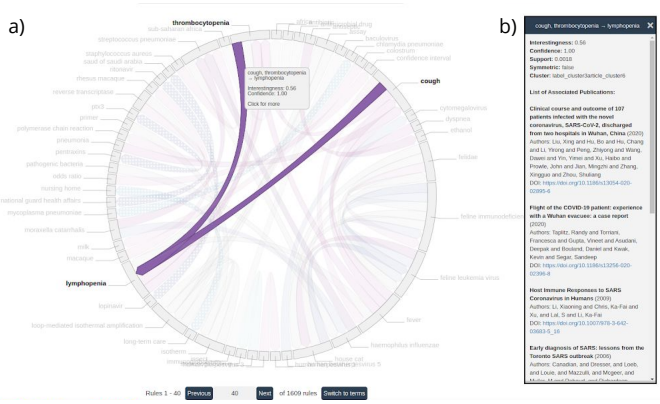

Fig. 7. Hovering the ribbons (left) in the circular view and nodes (right) in the graph view highlights the corresponding rule coloring the remaining in gray, while popping up a tooltip with details about the hovered rule and a button for recovering a list of publications related to the involved items.

Details-on-demand is provided through hovering over rules (i.e. ribbons or diamonds), which highlights the object and displays a tooltip that presents detailed information regarding the rule and the involved items, including of textual description of the measures *confidence, support, and interestingness*, and the name of corresponding the cluster (Fig. 7a). Furthermore, upon clicking on a rule, detailed information about the data source is displayed in an adjacent window (Fig. 7b). Since we are dealing with a dataset describing association rules within named entities identified in publications, the window lists the publications related to the named entities of the rule. We provide information on the title, authors, publication year, and the digital object identifier (DOI) of the publication. This feature allows the analyst to (1) consult the source of the visualized information, and (2) expand the description and understanding of association rules with contextual information.

E. Interactive synchronization

The three proposed views are independent but synchronized in a way that users may select subsets of data (rules or items) in one views to explore in another. The scatter plot embeds interactive features to enable the visual filtering of rules and their exploration on the circular paginated view of subsets. A context menu embedded in each diamond symbol allows users to select the set of rules represented to explore on the circular view. From the circular paginated view of subsets, the user may switch to the association graph view through a context menu embedded in each arc, which allow the user to choose to explore the rules involving that item when it is in the antecedent or consequent side of rules.

F. Interactive prompting

The system also includes an interactive prompting functionality to help users discover rules. In the association graph view, when the user searches for a NE in the search bar, the system displays all rules involving that NE and it prompts the user to rules involving every other NE pertaining the antecedent/consequent itemsets. Similarly, in the circular view, further to the rules within the selected NEs, the system displays rules which part of NEs pertain the selection. This feature enables the user to discover what other NEs are related to the searched item and to browse those "related" rules and NEs expanding the analysis.

V. Qualitative evaluation with Domain Experts

We conducted semi-structured interviews of about one hour via video-conference calls with two independent experts (over 15 years of experience), one from the the domains of data mining and another from biomedicine, who assessed the feasibility and usability of ARViz through the resolution of a set of domain-related tasks.

A. Material and Methods

Before starting the interview, the participants signed a Terms and Conditions agreement authorizing us to anonymously use their data in this study. The interview started with an overall exposition of the visual and interactive tools, and their objectives. The participant answered a standard socio-demographic questionnaire comprising standard profiling questions.

During a learning phase, the experimenter demonstrated the process of solving an analytic task using the interface, which the participant had to repeat afterwards with supervision of the experimenter. In the trial phase, the participants were asked to perform by themselves four sets of analytic tasks using the visualization interface. After completing each task the participants were asked to rate the difficulty of the tasks from 1 (very difficult) to 5 (very easy). During both phases, the participant would share their screen to allow the experimenter to follow their actions while performing the tasks. Comments uttered by the users were collected using thinking aloud protocol and duly recorded.

In the post-test phase, the participants answered a set of questions rating the importance of representing certain data information (e.g., measures of interest, source of data, etc) and providing certain interactive features (e.g., filtering, sorting, etc) in a 5-point Likert scale. We used the well-known System Usability Scale (SUS) [19] questionnaire to evaluate usability aspects. We also assessed the workload engendered by our tools through the Raw NASA-TLX (Task Load Index) [20] (RTLX) questionnaire, a multi-dimensional rating procedure that provides an overall workload score based on the average of ratings of six workload-related factors: mental demand, physical demand, temporal demand, own performance, effort, and frustration.

B. Results

Both participants solved all the tasks and assessed the solving process as easy or very easy. We observed that P1 solved all tasks only using the circular view, even though they seemed to understand the distinct roles of each view during the learning phase, while expressing to find particularly important the context menu in the circular view allowing to launch the association graph view for a particular item. P1 also never used the browsing tool to create subsets of items, only rules, which misled the answer of the following task: "Identify the three terms with the most association rules". By browsing rules without a particular order, P1 focused on the *displayed* items with the most rules, which are not necessarily the ones with the most associated rules in the dataset.

Both participants found very important to represent the connection between antecedent and consequent items, and the measures of interest. They also considered very important to filter the data according to items, clusters and measures of interest. They assessed as important to show related items to the searched one, but they do not find it an essential feature (score of 4/5). P2 found very important to show the clear relationship between items, the symmetry of rules, the source of data, and sorting rules and NEs, while P1 considered it important but not essential. Particularly, P1 mentioned that sorting rules and NEs would be necessary only for tasks such as "find the most confident or interesting rules", while being less useful to other types of tasks.

The overall RTLX scores showed that the perceived workload was slightly higher for P1 than P2, which scores were 55 and 46.7, respectively. P1 associated the physical demand to the action of positioning the cursor on the elements since, in P1's opinion, some of them were too small. P1 also assessed his own performance with higher score than P2, which might be associated to how they have previously assessed the success of their tasks. The SUS scores were 85 and 87.5 points for P1 and P2, respectively, which according to the Curved Grading Scale [21], indicates that they found the usability to be excellent. Both participants agreed that most people would not learn to use the system very quickly, which P2 believes to be due to the vocabulary we used in the interface (e.g., association rules, symmetry, etc). These concepts are particularly familiar to people working in the field of data mining, but not for people in other domains such as researchers in biomedicine, which are the second type of potential users of our visualization while applied to the COVID-19 scientific literature. P2 also found it difficult to understand the meaning of clusters and how to use them.

Both participants appreciated the link between the circular and association graph views, since that allows them to directly explore a NE of interest. Particularly, P2 appreciated the minimalist and organized design of the association graph and the interactive prompting, since that enable them to extend the analysis to NEs that they have not thought about before.

VI. Discussion

We presented ARViz[1], a visual approach designed to assist the rummaging process of association rules extracted from RDF knowledge graphs.

Analytical tasks. We propose three ways to approach the resolution of tasks based on association rules: the first one is an overview of all rules through a scatter plot (e.g., give the distribution of association rules according to confidence and interestingness); the second way is through a paginated exploration of subsets of rules using a chord diagram chart (e.g., identify the relationships between items or itemsets within a particular subset of rules or items); the third way is through an association graph that focus on itemsets (e.g., identify the rules associated to a particular NE). These views and tasks are complementary to allow a complete exploration the rules. The chord diagram and the association graph support the recovering of publications covering the named entities involved in a particular rule, which provides contextual information that allows to expand the analysis and to understand the relationship between named entities.

Visual design and interactions. We combined three complementary visualization techniques while improving the visual representation of attributes describing confidence, interestingness and symmetry of rules, which were only partially covered in previous works, specially in the visualization approach using a circular graph presented in [5]. We propose filtering criteria to allow the exploration of subsets of rules according to clusters of publications and named entities, or a selection of named entities of interest. We also provide an interactive prompting to assist the discovery of named entities within selected association rules by suggesting to the user related named entities to the one they searched.

Usability and suitability. The semi-structured interviews with two expert strongly suggest that our approach is an effective tool for assisting the exploration of association rules. The experts suggested that a pedagogic approach should be adopted to introduce the system to lay users in data mining, allowing them to apply the tool in their working routine to better understand and establish the relationship between named entities of interest. Nonetheless, the high degree of success in the execution of tasks and overall feeling that task as easy/very easy to perform with our tools, demonstrate that the approach

[1] Accessible at http://covid19.i3s.unice.fr:8080/arviz/

is feasible for solving domain-related tasks. Further studies are needed to evaluate the interface with a larger sample of expert users to confirm these hypotheses.

Generalization. Although this paper covers data describing the COVID-19 scientific literature, the visualization interface could handle any dataset of association rules regardless the underlying subject. Every feature can be equally used for other datasets and they can be easily adapted to encode different measures of interest. However, for data covering a completely different subject (e.g., products in a supermarket), the details-on-demand would have to be adapted offering further information (e.g., products description or supermarkets containing the selected products).

VII. Conclusions and future work

We presented an interactive visualization to assist the exploration of association rules. It was evaluated on rules extracted from a dataset describing the COVID-19 scientific literature. Our approach employs a scatter plot, a chord diagram, and an association graph to support overview of rules, and rule- and item-based analytic tasks. We represent attributes describing confidence, interestingness and symmetry of rules, and support the exploration of subsets of rules according to clustering results. The visualization of these attributes related to rules is original in the literature of data mining. We also provide an interactive prompting that aims to assist the discovery of named entities within selected association rules. Further, we allow the user to recover the publications covering the NEs involved in a particular association rule. A complementary approach such as ours for exploring association rules is totally novel in the literature.

Based on the feedback from expert users, we conclude that the results are encouraging, since participants from both data mining and biomedicine domains solved all proposed tasks with success and judged excellent the usability of the user interface. Future work includes the dissemination of the tool to researchers working with publication related to the COVID-19. In a longer run, we expect to redesign the visualization pipeline to support different RDF datasets and process them through the mining algorithm that extract the association rules to be visualized with our interface.

VIII. Acknowledgment

We acknowledge Inria for funding the CovidOnTheWeb project[2], in the context of which this research was conducted. We would also like to thank the participants of the experiment for allowing us to borrow their valuable time and knowledge.

References

[1] F. Gandon, "A Survey of the First 20 Years of Research on Semantic Web and Linked Data," *Revue des Sciences et Technologies de l'Information*, Dec. 2018.

[2] F. Michel, F. Gandon, V. Ah-Kane, A. Bobasheva, E. Cabrio, O. Corby, R. Gazzotti, S. Marro, T. Mayer, M. Simon, S. Villata, and M. Winckler, "Covid-on-the-web: Knowledge graph and services to advance covid-19 research," 09 2020.

[3] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," in *Semantic services, interoperability and web applications: emerging concepts*. IGI global, 2011, pp. 205–227.

[4] A. Unwin, H. Hofmann, and K. Bernt, "The twokey plot for multiple association rules control," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2001, pp. 472–483.

[5] C. P. Rainsford and J. F. Roddick, "Visualisation of temporal interval association rules," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2000, pp. 91–96.

[6] L. Yang, "Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates," in *International Conference on Computational Science and Its Applications*. Springer, 2003, pp. 21–30.

[7] D. Bruzzese and P. Buono, "Combining visual techniques for association rules exploration," in *Proceedings of the working conference on Advanced visual interfaces*, 2004, pp. 381–384.

[8] Y. A. Sekhavat and O. Hoeber, "Visualizing association rules using linked matrix, graph, and detail views," 2013.

[9] K. Techapichetvanich and A. Datta, "Visar: a new technique for visualizing mined association rules," in *International Conference on Advanced Data Mining and Applications*. Springer, 2005, pp. 88–95.

[10] C.-W. Cheng, Y. Sha, and M. D. Wang, "Intervisar: An interactive visualization for association rule search," in *Proc. 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2016, pp. 175–184.

[11] K. huat Ong, K. leong Ong, W.-K. Ng, and E.-P. Lim, "Crystalclear: Active visualization of association rules," in *In ICDM'02 International Workshop on Active Mining AM2002*. Press, 2002.

[12] M. Hahsler, "arulesviz: Interactive visualization of association rules with r." *R Journal*, vol. 9, no. 2, 2017.

[13] C. H. Yamamoto, M. C. F. de Oliveira, and S. O. Rezende, "Visualization to assist the generation and exploration of association rules," in *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*. IGI, 2009, pp. 224–245.

[14] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "CORD-19: The COVID-19 Open Research Dataset," *arXiv e-prints*, p. arXiv:2004.10706, 2020.

[15] M. A. Musen, N. F. Noy, N. H. Shah, P. L. Whetzel, C. G. Chute, M.-A. Story, B. Smith, and N. team, "The national center for biomedical ontology," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 190–195, 2012.

[16] L. Foppiano and L. Romary, "entity-fishing: a DARIAH entity recognition and disambiguation service," in *Digital Scholarship in the Humanities* , Tokyo, Japan, Sep. 2018.

[17] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, "Improving efficiency and accuracy in multilingual entity extraction," in *Proc. 9th International Conference on Semantic Systems*, 2013, pp. 121–124.

[18] L. Cadorel and A. G. B. Tettamanzi, "Mining RDF Data of COVID-19 Scientific Literature for Interesting Association Rules," in *International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'20)*, 2020, p. to appear.

[19] J. Brooke, "Sus: a 'quick and dirty' usability," *Usability evaluation in industry*, p. 189, 1996.

[20] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[21] J. Sauro and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.

[2] https://www.inria.fr/en/covid-web