Interactive redescription set mining and exploration

Iva Kozjak* and Matej Mihelčić*
* University of Zagreb, Faculty of Science/Department of Mathematics, Zagreb, Croatia
kozjak.iva@gmail.com, matmih@math.hr

***Abstract*—InterSet is a client-server web application that allows targeted and contextual redescription set exploration. The main drawback of this tool is that it uses fixed, precomputed sets of redescriptions to allow obtaining novel knowledge. In this work, we significantly extend the capabilities of InterSet by adding the possibility to create new redescriptions on predefined data. The main advantage of this is that the user can create new redescriptions in any step of exploration depending on the context and current research hypothesis, utilizing entity or attribute constraint-based redescription mining and enriching the existing set of redescriptions with newly obtained knowledge. This procedure further enhances the potential of forming and exploring new research hypotheses using large sets of redescriptions. The redescription mining ability of the InterSet tool is achieved by utilizing the CLUS-RM algorithm. Usefulness of the obtained environment for interactive redescription set creation, targeted and contextual exploration is demonstrated on real-world use case datasets.**

Keywords*—*InterSet, interactive redescription mining, redescription set exploration, world countries, dblp authorship network, phenotypes

I. INTRODUCTION

Redescription mining [1] is a field of data mining that has two main goals: a) discover subsets of entities in the data that can be described in multiple ways, b) construct appropriate redescriptions of discovered subsets of entities. Finding multiple descriptions (called redescriptions) of the same or very similar subsets of entities, in an unsupervised manner, is the main distinguishing factor of this approach compared to the related fields of clustering [2], [3], conceptual clustering [4], [5], rule learning [6] and subgroup discovery [7]–[9]. In contrast to the association rule mining [10]–[12], where discovered rules form implications, descriptions obtained by redescription mining are in strong equivalence relations.

Redescription mining approaches either return all redescriptions that satisfy some predefined quality criteria (e.g. [1], [13]–[15]) or use redescription set optimization to obtain a smaller set of diverse redescriptions that offer favourable trade-off between many different quality criteria (e.g. [16], [17]). Approaches that utilize filtering, often create a large set of results. This set is very hard to explore and extremely difficult to reason about without the use of additional tools.

There exist two main tools that allow obtaining additional knowledge from sets of redescriptions: the Siren [18] and the InterSet [19]. The Siren is an interactive redescription mining framework that allows mining redescriptions, inspecting and expending individual redescriptions, filtering, embedding instances described by a redescription into some lower-dimensional plane etc. InterSet [19] allows targeted and conceptual redescription set exploration. The main idea was to allow exploration of large sets of redescriptions by allowing redescription subset creation and selection, based on entities described by redescriptions, attributes contained in their queries and general redescription quality measures. The tool also allows statistical analyses of individual redescriptions and relating different redescriptions. The main drawback of the InterSet tool, prior to this work, was that it worked with the precomputed sets of redescriptions.

In this work, we present a way to remove the aforementioned disadvantage of the InterSet by incorporating interactive and constrained-based redescription mining into this tool. The resulting tool, among others, has the ability to perform interactive redescription mining, potentially using entity, attribute-based constraints or the combination thereof and to allow interactive targeted and contextual redescription set exploration.

II. DEFINITIONS AND NOTATION

The input data $\mathcal{D}$ is in a form of one or more tables $D_i,\ i = 1, \ldots, n$, where each table contains values of Boolean, categorical or numerical attributes for one, shared set of entities. We denote a set of all attributes with $\mathcal{A}$ and a set of entities with E. Data tables D_i form views of the input data $\mathcal{D}$. $\mathcal{W} : \mathcal{A} \mapsto \mathbb{N}$ is a mapping that returns the view of some attribute $a \in \mathcal{A}$. Views can, for example, represent repeated experiments, experiments performed in different conditions or by different institutions, these can be natural groupings of attributes based on domain knowledge or the current research problem.

Knowledge obtainable from the data is represented in a form of logical formulae, where each formula consists of variables (attributes with corresponding conditions) and logical connectives (conjunction, disjunction and negation). Such formulae are called queries (q). The choice of allowed logical operators and the definition of rules of generative grammar define the set of allowed queries (Q), also called a query language. A set of entities described by a query q (satisfying the logical formula) form the support set of this query $supp(q)$. Redescriptions are tuples of queries $R = (q_1, q_2, \ldots q_n)$ where $supp(q_1) \sim$

$supp(q_2) \sim \cdots \sim supp(q_n)$ (queries forming redescriptions should have similar support sets). The set of all attributes contained within queries of a redescription R is denoted by $attrs(R)$. Similarity of support sets of queries is measured using the Jaccard index [20]. The corresponding similarity score $J(R) = (|\cap_{i=1}^{n} supp(q_i)|)/(|\cup_{i=1}^{n} supp(q_i)|)$ represents the accuracy of a redescription R. Statistical significance of a redescription is measured as: $p(R) = \sum_{i=|supp(R)|}^{|E|} \binom{|E|}{i} \prod_{j=1}^{n}(|supp(p_j)|/|E|) \cdot (1 - \prod_{j=1}^{n}(|supp(p_j)|/|E|))$. The complexity of a redescription is measured as $comp(R) = min((|attrs(R)|/k), 1)$, where k represents the maximal desired size of $|attrs(R)|$.

A set of redescriptions is denoted by $\mathcal{R}$. The average redundancy of a redescription R, contained in a set of redescriptions $\mathcal{R}$, based on attributes contained in its queries is denoted as: $AAJ(R, \mathcal{R}) = (\sum_{R_i \in \mathcal{R},\ R_i \neq R} J(attrs(R), attrs(R_i)))/(|\mathcal{R}| - 1)$. The average redundancy of a redescription based on entities it describes is denoted as $AEJ(R, \mathcal{R}) = (\sum_{R_i \in \mathcal{R},\ R_i \neq R} J(supp(R), supp(R_i)))/(|\mathcal{R}| - 1)$.

We say that a redescription R is associated with the entity cluster $\mathcal{G}$ if $J(supp(R), G) \geq \delta$, where δ denotes the required similarity level. The average homogeneity of an entity cluster $\mathcal{G}$ is defined as $avHom(\mathcal{G}) = \sum_{R \in \mathcal{G}} J(supp(R), \mathcal{G})/|\mathcal{G}|$.

Given a set of user-defined constraints on redescriptions $\mathcal{C}$, the input data $\mathcal{D}$ and the query language Q, the goal of redescription mining is to discover all redescriptions that satisfy constraints defined in a set $\mathcal{C}$.

III. Enabling Interactive Redescription Mining with the InterSet

InterSet is a web application, where the client part was created using standard technologies: `HTML`, `CSS`, `JavaScript` and the `angular.js`. The server part of the InterSet tool was originally implemented using the `node.js` environment (see Figure 1).

In this work, we replace the `node.js` backend with the `Java Spring Boot`. This allows stronger integration of the CLUS-RM algorithm [17], extended with the ability to perform constraint-based redescription mining [21], implemented in the `Java` programming language, with the InterSet tool [19].

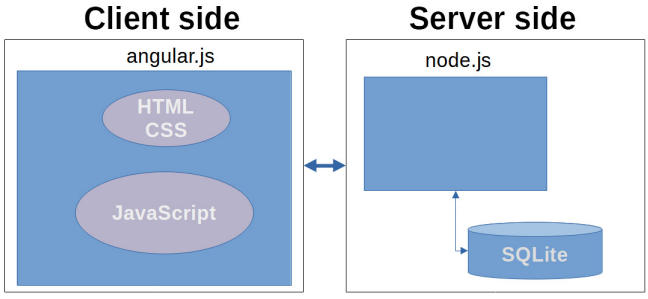

Fig. 1: Architecture of the InterSet tool prior to the proposed extensions.

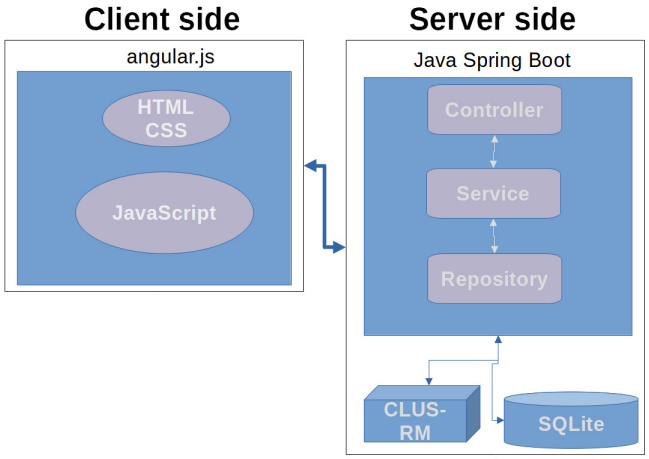

Fig. 2: Architecture of the InterSet tool after the proposed extensions.

The newly obtained structure of the InterSet can be seen in Figure 2. The *controller* package of the InterSet tool contains one class per dataset: the *DBLPRedescriptionsController*, *PhenotypeRedescriptionsController* and *TradeRedescriptions-Controller*. The *service* package contains a class *RedescriptionsService* and the *repository* package contains a class *RedescriptionsRepository*.

The CLUS-RM algorithm [21] has been adopted for integration with the InterSet tool. The main modification is to allow adding joined entity-attribute constraints into the constrained-based redescription mining process. The constraints have been modified to contain three main categories: *none* (no constraints, perform regular redescription mining), *soft* (at least one entity and at least one attribute from the predefined set of constraints must be contained in the support and attribute sets of the newly created redescriptions), *hard* (all entities and attributes from the predefined set of constraints must be contained in the support and attribute sets of the newly created redescriptions).

The InterSet tool [19] contains three main exploration views: the Entity view, the Attribute view and the Redescription quality view. The first two are modified to allow interactive redescription mining, but the newly added redescriptions in any of these views cause updates of the statistics reported in the Redescription quality view.

A. Newly added functionality to the InterSet

When the user selects one cluster from the SOM (Self-organizing map) [22], contained in the *Entity - based exploration* tab, a table of entities contained in this cluster is displayed. The main goal of enabling the use of constrained-based redescription mining is to allow discovering new redescriptions that describe all or large subsets of entities contained in the selected cluster. This allows expanding the knowledge base and can lead to proving, disproving or forming new research hypotheses.

The most important settings of the CLUS-RM algorithm can be adjusted through the user interface, and

the required constraints and type can be selected using corresponding radio-buttons. When both the SOM cluster and the constraint type are selected, the CLUS-RM algorithm can be invoked. By clicking the run button, the selected properties are sent to the server side of the application where the CLUS-RM algorithm is executed, and the result file is created. For each of the resulting redescriptions R, the Jaccard similarity index of entities with the selected cluster G is being calculated as $J(supp(R), G) = (|supp(R) \cap E|)/(|supp(R) \cup E|)$. Redescriptions and the aforementioned Jaccard similarity are saved and sent to the client side, where they are displayed in a table. This table allows determining the support sets of the newly created redescriptions, which redescriptions describe entities from the selected cluster in the most suitable manner, and it provides information about the most similar redescription, to each of the newly created redescriptions, already contained in the database. Based on the information provided in this table, the user can save appropriate redescriptions into the database.

In the *Attribute - based exploration* tab, the user can select pairs of attributes from the heatmap that visualizes statistics about frequency and co-occurrence of attributes contained in the queries of redescriptions from the database. The main advantage of incorporating constraint-based redescription mining in this view is to allow discovering new redescriptions containing interesting set of attributes in their queries. This allows confirming, exploring, expanding or forming hypothesis about associations of attributes in the dataset.

Similarly as in the *Entity - based exploration* tab, depending on the choice from the heatmap, the appropriate set of attributes is saved and displayed as a potential set of attribute constraints. The appropriate radio-buttons allow selecting constraint type, and the CLUS-RM algorithm can be invoked using the run button. Again, a table is created containing all newly created redescriptions with information about the most similar redescription already contained in the database. An additional table provides description of the attributes contained in the queries of a selected newly discovered redescription. Based on the provided information, the user can save appropriate redescriptions in the database.

Running the CLUS-RM algorithm with conditions on both entities and attributes is possible if the selection (set of entities or attributes) is made and saved in one of the views and then loaded in the other view. In that case, the set of selected attributes is displayed alongside entities from the selected cluster.

B. Saving newly created redescriptions into the database

The user can save subsets of newly found redescriptions from the table by marking the appropriate checkboxes in the table. The save action causes updates of multiple tables in the database, necessitates computation of redescription set based quality scores (AEJ,AAJ) of all newly added redescriptions and re-computation of these quality scores for all redescriptions previously contained in the database.

For a set of redescriptions $\mathcal{R}_{db}$ and a set of redescriptions $\mathcal{R}_{new}$, the AEJ and AAJ for redescriptions contained in $\mathcal{R}_{new}$ are computed as:
$AEJ(R, \mathcal{R}_{db} \cup \mathcal{R}_{new}) = (\sum_{R_i \in \mathcal{R}_{db} \cup \mathcal{R}_{new},\ R_i \neq R} J(supp(R), supp(R_i)))/(|\mathcal{R}_{db} \cup \mathcal{R}_{new}| - 1)$.
$AAJ(R, \mathcal{R}_{db} \cup \mathcal{R}_{new}) = (\sum_{R_i \in \mathcal{R}_{db} \cup \mathcal{R}_{new},\ R_i \neq R} J(attrs(R), attrs(R_i)))/(|\mathcal{R}_{db} \cup \mathcal{R}_{new}| - 1)$. The updates for redescriptions contained in $\mathcal{R}_{db}$ are computed as:
$AEJ(R, \mathcal{R}_{db} \cup \mathcal{R}_{new}) = (AEJ(R, \mathcal{R}_{db}) \cdot |\mathcal{R}_{db} - 1| + \sum_{R_i \in \mathcal{R}_{new}} J(supp(R), supp(R_i)))/(|\mathcal{R}_{db} \cup \mathcal{R}_{new}| - 1)$.
$AAJ(R, \mathcal{R}_{db} \cup \mathcal{R}_{new}) = (AAJ(R, \mathcal{R}_{db}) \cdot |\mathcal{R}_{db} - 1| + \sum_{R_i \in \mathcal{R}_{new}} J(attrs(R), attrs(R_i)))/(|\mathcal{R}_{db} \cup \mathcal{R}_{new}| - 1)$.

Complete code of the current version of the tool and links to the older versions of the InterSet tool and the CLUS-RM algorithm can be found on: https://github.com/ivakozjak/MIPRO2023.

IV. Applying InterSet for scientific discovery

We apply InterSet to the three well-known use-case datasets: the Country dataset, the Phenotype dataset and the DBLP dataset (see [19] for more details). For all presented redescriptions $p(R) \leq 0.01$.

A. Discovery on the Country dataset

We perform constraint-based redescription mining with entity constraints on 3 different clusters from the SOM map on the Country dataset, as demonstrated in Figure 3.

The top left cluster has the highest homogeneity 0.3374 and 2737 redescriptions describing at least one Country contained in this cluster. This cluster contains 11 highly developed Western European countries that share many common trading and socio-demographic patterns. Using the presented extensions of the InterSet, we managed to find redescription R_1 (first in Table I) with entity Jaccard of 0.91 with this cluster. After adding this redescription to the database, the overall homogeneity of the top left cluster increased to 0.3376. The newly discovered redescription describes that small percentage of population of these countries was employed in agriculture, there was high trade in Stocks and that these countries exhibited significantly larger export than import of specialized machinery and had high import of manufactured goods. Although there already exists redescriptions with equal support set in the database, the newly discovered redescription has an attribute redundancy of 17% with the most similar redescription, thus it provides novel knowledge about the entities contained in the first cluster.

The neighbouring cluster (first in the second row) mostly contains Central and Eastern European countries. This cluster originally had homogeneity of 0.2163 and there existed 3002 redescriptions describing at least one country from this cluster. The newly discovered redescription R_2 (second in Table I) with entity Jaccard of 0.72 with the aforementioned cluster when added to the database

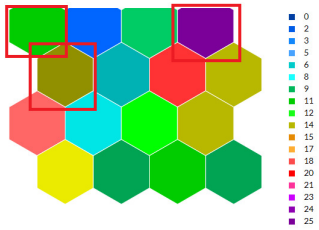

Fig. 3: Studied clusters on the Country dataset.

increases the overall homogeneity of this cluster to 0.2165. The most important features described by R_2 include higher unemployment of females ($2.5 - 18\%$), larger percentage of bad employment ($5 - 35\%$), import of non-metallic manufactures and labour-intensive manufactures, pronounced export of metal manufactures and mixed ratio of export to import of fuels. The newly discovered redescription describing the second selected cluster has a maximal redundancy of 90% with a redescription already contained in the database, but the attribute redundancy of only 9% with a redescription having most similar support.

The third selected cluster contains a diverse set of 28 countries (e.g. Afghanistan, Algeria, Angola, Cameron, Congo, Iraq, Mali, Niger, Tajikistan, Yemen). This cluster originally had homogeneity of 0.2396 and had 2019 redescriptions describing at least one country from this set. For this cluster, we discovered 7 redescriptions with entity Jaccard ≥ 0.4 with the selected cluster. After adding all 7 discovered redescriptions, the homogeneity of this cluster increased to 0.2403. Redescription R_3 (third in Table I) with entity Jaccard of 0.48 with the selected cluster is the most related redescription obtained in our analyses. This redescription describes 27 countries from the selected cluster and 18 countries from the neighbouring 3 clusters. Thus, it contains relevant information for many related countries. The main characteristics of these countries are relatively small percentage of population aged 64 or above ($2.2 - 5.0\%$), high percentage of rural population ($22.1 - 80\%$) and higher import than export of medicinal and pharmaceutical products. Level of export of beverages and plastics in non-primary form, level of export to import ratio of footwear and plastics in primary form and export to import ratio of fuels varies between countries. Newly discovered redescriptions associated with the third selected cluster have a maximal entity redundancy with other redescriptions contained in the database of $58 - 69\%$ and attribute redundancy of $0 - 17\%$ with redescriptions having the most similar support. Thus, newly discovered redescriptions contain useful knowledge, previously unavailable in the database.

TABLE I: Redescription examples obtained on the Country dataset.

q_1	q_2	J	$\lvert supp \rvert$
$1.2 \leq E_{AG} \leq 4.9 \wedge$ $27.0 \leq ST \leq 166.6$	$1.1 \leq E/I_{SM} \leq 4.3$ $\wedge\ 58 \leq I_{MG} \leq 78$	0.67	12
$2.5 \leq U_F \leq 28.0 \wedge$ $5.0 \leq E_{bad} \leq 32.1$	$1 \leq I_{NMM} \leq 2 \wedge$ $0.1 \leq E/I_F \leq 1.2 \wedge$ $8 \leq I_{LIRIM} \leq 16 \wedge$ $2 \leq E_{MM} \leq 6 \wedge$	0.53	18
$2.2 \leq P_{64} \leq 5.0 \wedge$ $22.1 \leq R_P \leq 88.8$	$0 \leq E/I_{MP} \leq 0.19 \wedge$ $0 \leq E_B \leq 2 \wedge$ $0 \leq E/I_{FW} \leq 9.2 \wedge$ $0 \leq E_{Pl} \leq 1 \wedge$ $0 \leq E/I_F \leq 161.2 \wedge$ $0 \leq I_{PPSN} \leq 1 \wedge$ $0 \leq E/I_{PPF} \leq 9.2 \wedge$	0.54	55
$0 \leq C_{Cv} \leq 53.5 \wedge$ $2.2 \leq P_{64} \leq 11 \wedge$ $50 \leq P_{15,64} \leq 73$ $\wedge\ 18 \leq P_{14} \leq 47.6$	$0 \leq E/I_{SM} \leq 0.64$	0.63	103
$10.6 \leq P_{64} \leq 24.4$	$0.65 \leq E/I_{SM} \leq 61.9$	0.5	29
$0.3 \leq W_R \leq 1.9 \wedge$ $59 \leq L_M \leq 67.8 \wedge$ $30 \leq E_{IM} \leq 50 \wedge$ $11 \leq ST \leq 167 \wedge$ $13 \leq C_{Cv} \leq 100 \wedge$ $-0.2 \leq P_G \leq 0.5$ $\wedge\ 22.4 \leq R_P \leq 47.2$	$1 \leq E/I_{MT} \leq 1.8 \wedge$ $2 \leq E_{MM} \leq 5 \wedge$ $0.18 \leq E/I_{DP} \leq 3.1$ $\wedge\ 0 \leq E/I_{CR} \leq 0.95$	0.75	9

We test two attribute-based constraints. First, we search redescriptions with hard constraints on the most-frequently co-occuring attributes P_{64}, E/I_{SM}. Using these constraints, we managed to obtain new redescriptions containing both attributes in their queries. Redescriptions R_4 and R_5 (fourth and fifth in Table I) are examples of such redescriptions. Both redescriptions have very small attribute redundancy ($1 - 4\%$) with respect to used attributes to redescriptions with the most similar support contained in the database. Thus, they provide novel associations of the selected pair of attributes.

Soft attribute constraints C_{Cv}, E/I_{MP} allow discovering redescriptions that contain either of these attributes in their queries. Redescription R_6 (sixth redescription in Table I), discovered using soft constraints on the aforementioned attributes, contains C_{Cv} but not E/I_{MP}.

B. Discovery on the Phenotype dataset

Using the proposed extensions of the InterSet tool, we discovered a redescription R_7 (first in Table II) with entity Jaccard of 0.97 with the first (top left) cluster in the SOM map. Although there existed a redescription with equal support set in the database, the newly discovered redescription has an attribute redundancy of 10% with this redescription, providing novel information about the target subset of bacteria. This redescription describes unicellular photosystem bacteria residing in the water. After adding this redescription to the database, the homogeneity of the first cluster increased from 0.5160 to 0.5163. Redescrip-

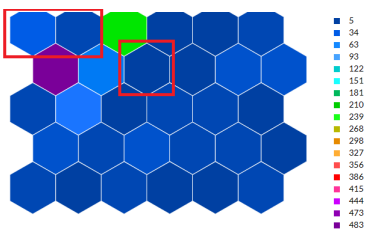

Fig. 4: Studied clusters on the Phenotype dataset.

tion R_8 (second in Table II) has the entity Jaccard of 0.79 with the second selected cluster in the first row.

TABLE II: Redescription examples obtained on the Phenotype dataset.

q_1	q_2	J	$\|supp\|$
$unicPhotosyOF$	$C_{260} \wedge C_{349} \wedge C_{321} \wedge C_{1624} \wedge C_{5474} \wedge C_{1530} \wedge C_{99}$	0.84	32
$hlHsArSFNaCl \wedge \neg TRHypTP \wedge \neg GSP$	$C_{4749} \wedge C_{442} \wedge \neg C_{3890} \wedge \neg C_{1481} \wedge \neg C_{4412} \wedge \neg C_{1753}$	0.61	17
$cytFeHCDis \wedge \neg unicPhotosysOF \neg OF \wedge \neg TPHypTP \wedge \neg GSP \wedge oxSRA$	$(\neg C_{3102} \wedge \neg C_{4129} \wedge \neg C_{3764} \wedge \neg C_{3373} \wedge \neg C_{419} \wedge C_{442} \wedge C_{3012} \wedge C_{2033}) \vee (\neg C_{3580} \wedge \neg C_{3611} \wedge \neg C_{1205} \wedge \neg C_{2926} \wedge \neg C_{1844} \wedge C_{1183} \wedge C_{4881} \wedge C_{1379})$	0.67	12
$oxSA \wedge shCoc \wedge \neg metSulf \wedge \neg metMeth \wedge hypAHdrVHet$	$C_{3354} \wedge C_{4025}$	0.67	10
$brChUlMyEnd \wedge \neg khabSl \wedge \neg ecsyEnv$	$C_{1853} \wedge C_{5468} \wedge \neg C_{123} \wedge \neg C_{4286} \wedge \neg C_{3373}$	0.64	7

Although there exists a redescription with very similar support set with this redescription (entity Jaccard of 0.94), their attributes are different, thus we obtained novel knowledge about bacteria contained in the target cluster. Bacteria described by this redescription live in extreme (hypersaline) environments. The homogeneity of the second cluster increased from 0.7130 to 0.7136 after adding the newly discovered cluster to the database. Given that only 127 redescriptions were originally associated with this cluster, the importance of the newly discovered redescription is substantial. Redescription R_9 has entity Jaccard of 0.54 with the third cluster (third in the second row). This redescription describes all bacteria from the target cluster except Geobacter uraniireducens. Additionally, it describes 5 bacterial species from the genus Desulfovibrio contained in the neighbouring cluster (third row, fourth cluster). Given that the database contains only 41 redescriptions that describe at least one species of the bacteria contained in the third cluster and its homogeneity is 0.41, the discovered redescription provides a very useful additional knowledge. Addition of the presented redescription increases homogeneity of the third cluster to 0.413.

We performed constrained-based redescription mining with soft attribute constraints $shCoc$, C_{3373} and present two obtained redescriptions R_{10} and R_{11} (fourth and fifth redescription in Table II). Redescription R_{10} contains positive information about bacteria that are Coccus shaped whereas the second redescription contains information about a set of bacteria that does not contain COG_{3373}. The described set of bacteria is connected to the terms buruli ulcer and chancroid.

C. Discovery on the DBLP dataset

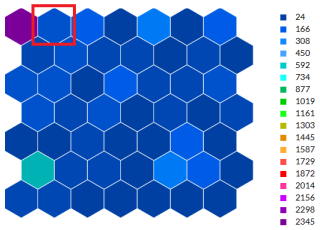

Fig. 5: Studied clusters on the DBLP dataset.

On the DBLP dataset, we focus on the second cluster in the first row. This cluster contains 109 different authors of scientific manuscripts, where 144 redescriptions describe at least one author from the cluster. Redescriptions associated with this cluster are very general, mostly containing negations of attributes (saying what is not true for many authors contained in their support sets). As a result, the homogeneity of this cluster is very small 0.0213. Using authors contained in this cluster as soft constraints, we discovered 17 new redescriptions with entity Jaccard $\geq$ 0.03 with the selected cluster. The important aspect of the discovery is that newly obtained redescriptions contain some positive attributes, revealing conferences in which a part of the authors contained in the cluster published their manuscripts and their co-authors. After adding newly discovered redescriptions to the database, the homogeneity of this cluster increased to 0.0243. Redescription R_{12} presented in Table III has the largest entity Jaccard (0.0365) with the selected cluster. It describes a group of authors that co-authored their manuscripts with either Ravishankar K. Iyer, Amin Vahdat, Robbert van Renesse, Robert J. Stroud, E. N. Elnozahy, Michele Garetto, Michel Raynal or Maria L. Gini. The described authors published manuscripts in the ICDS (International Conference on Digital Society) and the Symposium on Reliable Distributed Systems. R_{12} describes 5 authors from the selected cluster, revealing information about their scientific activity and 15 authors from the neighbouring cluster (second in the second row).

V. Conclusions

We presented the extension of the InterSet tool that allows performing interactive and constrained-based redescription mining from the user interface of the InterSet tool. This important extension transforms InterSet

TABLE III: Redescription examples obtained on the DBLP dataset.

q_1	q_2	J	$\|supp\|$
$ICDS \wedge SRDS$	$R.K.Iyer \vee A.Vahdat \vee R.vanRenesse \vee R.J.Stroud \vee E.N.Elnozahy \vee M.Garetto \vee M.Raynal \vee M.L.Gini$	0.17	33

into targeted and contextual redescription mining and redescription set exploration tool.

Scientific merit of the approach is demonstrated on three different use-case datasets. The authors obtained novel redescriptions, not previously contained in the database, with significantly higher entity Jaccard index with the targeted SOM cluster than the homogeneity score of this cluster on all three datasets and demonstrated meaningfulness of the created redescriptions for the target cluster and the neighbouring clusters. Analyses were performed both for clusters with high and with low homogeneity score. As was expected, clusters with low homogeneity score contain heterogeneous entities, thus newly created redescriptions, obtained using constraint-based redescription mining, describe only subsets of the targeted clusters, potentially including neighbouring related entities. The authors also demonstrate functionality of constrained-based redescription mining with attribute-based constraints, where novel previously unknown associations were discovered for the targeted attributes both using hard and soft constraints. It should be noted that obtaining novel results is not always possible (e.g. with to strict constraints). In such cases, it is necessary to relax the constraints and re-run constraint-based redescription mining.

The obtained tool is one of only two interface-based interactive redescription mining tools and the only tool capable of incorporating information about redescription sets into constrained-based redescription mining and targeted and contextual redescription set exploration.

REFERENCES

[1] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R. F. Helm, "Turning CARTwheels: an alternating algorithm for mining redescriptions," in *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), Seattle, Washington, USA*. ACM, 2004, pp. 266–275.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comp. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[3] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *An. Dat. Sci.*, vol. 2, no. 2, pp. 165–193, 2015.

[4] R. S. Michalski, "Knowledge acquisition through conceptual clustering: A theoretical framework and an algorithm for partitioning data into conjunctive concepts," *J. Pol. Anal. Inf. Syst.*, vol. 4, no. 3, pp. 219–244, September 1980.

[5] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Mach. Learn.*, vol. 2, no. 2, pp. 139–172, 1987.

[6] J. Fürnkranz, D. Gamberger, and N. Lavrac, *Foundations of Rule Learning*, ser. Cognitive Technologies. Springer, 2012.

[7] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," in *Procedings of first European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'97), Trondheim, Norway*, ser. Lecture Notes in Computer Science, vol. 1263. Springer, 1997, pp. 78–87.

[8] N. Lavrač, B. Kavšek, P. A. Flach, and L. Todorovski, "Subgroup discovery with CN2-SD," *J. Mach. Learn. Res.*, vol. 5, pp. 153–188, 2004.

[9] F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus, "An overview on subgroup discovery: foundations and applications," *Know. Inf. Syst.*, vol. 29, no. 3, pp. 495–525, 2011.

[10] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in *Adv. Know. Disc. Dat. Min.* AAAI/MIT Press, 1996, pp. 307–328.

[11] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining - A general survey and comparison," *SIGKDD Expl.*, vol. 2, no. 1, pp. 58–64, 2000.

[12] M. Zhang and C. He, *Survey on Association Rules Mining Algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 111–118.

[13] A. Gallo, P. Miettinen, and H. Mannila, "Finding subgroups having several descriptions: Algorithms for redescription mining," in *Proceedings of the SIAM International Conference on Data Mining, (SDM'08), Atlanta, Georgia, USA*. SIAM, 2008, pp. 334–345.

[14] E. Galbrun and P. Miettinen, "From black and white to full color: extending redescription mining outside the boolean world," *Stat. Anal. Dat. Min.*, vol. 5, no. 4, pp. 284–303, 2012.

[15] T. Zinchenko, E. Galbrun, and P. Miettinen, "Mining predictive redescriptions with trees," in *IEEE International Conference on Data Mining Workshop, (ICDMW'15), Atlantic City, NJ, USA*. IEEE Computer Society, 2015, pp. 1672–1675.

[16] M. Mihelčić, S. Džeroski, N. Lavrač, and T. Šmuc, "A framework for redescription set construction," *Expert Syst. Appl.*, vol. 68, pp. 196–215, 2017.

[17] ——, "Redescription mining augmented with random forest of multi-target predictive clustering trees," *J. Intell. Inf. Syst.*, vol. 50, no. 1, pp. 63–96, 2018.

[18] E. Galbrun and P. Miettinen, "Mining redescriptions with siren," *ACM Trans. Knowl. Discov. Data*, vol. 12, no. 1, jan 2018.

[19] M. Mihelčić and T. Šmuc, "Targeted and contextual redescription set exploration," *Mach. Learn.*, vol. 107, no. 11, pp. 1809–1846, 2018.

[20] P. Jaccard, "The distribution of the flora of the alpine zone," in *New Phytologist*, vol. 11, 1912, pp. 37–50.

[21] M. Mihelčić, G. Šimić, M. Babić Leko, N. Lavrač, S. Džeroski, T. Šmuc, and for the Alzheimer's Disease Neuroimaging Initiative, "Using redescription mining to relate clinical and biological characteristics of cognitively impaired and alzheimer's disease patients," *PLOS ONE*, vol. 12, no. 10, pp. 1–35, 10 2017.

[22] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, Jan. 1982.