Redescription Mining: An Overview.

Esther Galbrun, Pauli Miettinen

Redescription Mining: An Overview

Esther Galbrun and Pauli Miettinen

***Abstract*—In many real-world data analysis tasks, we have different types of data over the same objects or entities, perhaps because the data originate from distinct sources or are based on different terminologies. In order to understand such data, an intuitive approach is to identify the correspondences that exist between these different aspects. This is the motivating principle behind *redescription mining*, a data analysis task that aims at finding distinct common characterizations of the same objects. This paper provides a short overview of redescription mining; what it is and how it is connected to other data analysis methods; the basic principles behind current algorithms for redescription mining; and examples and applications of redescription mining for real-world data analysis problems.**

***Index Terms*—Redescription mining, alternative characterizations, visualizations, data mining.**

I. Introduction

CONSIDER an ecologist who wants to understand the bioclimatic conditions that define species' habitats.[1] She has data on the regions where the species live and on the bioclimatic conditions (e.g. monthly average temperatures and precipitation) of those regions, and she would like to find explanations such as the following.

> *The areas inhabited by either the Eurasian lynx or the Canada lynx are approximately the same areas as those where the maximum March temperature ranges from* $-24.4\,°\mathrm{C}$ *to* $3.4\,°\mathrm{C}$.

The above is an example of a *redescription*. It *describes* regions of the earth in two different ways; on the one hand, by the fact that certain species inhabit them, and on the other hand, by the fact that they have a certain climate. We can see the areas described above in Figure 1. The medium purple colour denotes the areas where both of the above conditions hold (inhabited by one of the lynx species and with maximum March temperatures in the correct range), light red denotes the areas inhabited by one of the lynx species but where March temperatures are out of the range, and dark blue denotes the areas where the maximum March temperature is in the correct range but neither of the lynxes is found.

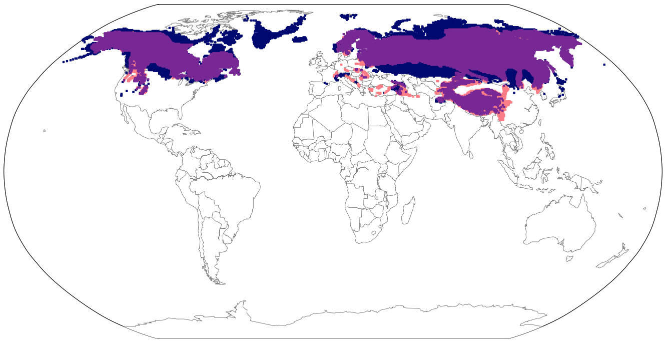

Figure 1. Map of a bioclimatic niche. The areas inhabited by either the Eurasian lynx or the Canada lynx (light red and medium purple) and the areas where the maximum March temperature is between $-24.4\,°\mathrm{C}$ and $3.4\,°\mathrm{C}$ (dark blue and medium purple).

Informally, a redescription is a pair of descriptions, both describing roughly the same entities (here, geographical regions). And, as we can see from this example, both the descriptions and what they describe can be of interest. The ecologist is interested in the descriptions in order to understand the *model* of the niche and in the geographical areas in order to understand where the niche is (or is not). While such redescriptions could be constructed manually, the goal of *redescription mining* is to find them automatically without any information other than the raw data (and some user-provided constraints). For instance, the ecologist should not have to define the species she is interested in. Rather, the goal of redescription mining is to find all redescriptions that characterize sufficiently similar sets of entities and adhere to some simple constraints regarding, for example, their type and complexity and how many entities they cover.

In this article, we present a brief overview of redescription mining. We start by giving the formal definition of the task in the next section. In Section III, we explain the main algorithmic ideas used in redescription mining, before discussing the techniques for removing redundant redescriptions, in Section IV. Sections V, VI, and VII contain, respectively, a brief study of the existing redescription mining tools, an outline of some example applications, and a summary of related methods. We present some open problems and directions to future work in the concluding Section VIII. We will not delve into the details of the different algorithms, tools, or applications. Such details can be found in the original publications, as well as in our recent tutorials[2] and book [1].

II. Formalizing the Task

Redescription mining can, of course, be applied to other use cases than bioclimatic niche finding, but we will use that example as our running example throughout this article. In this section we provide the formal definition of redescription mining. Our definition uses the so-called table-based model [1]; other, more general, formulations exist (see [1]), but that generality is unnecessary for the discussion in this article.

E. Galbrun is with Aalto University, Finland. email: esther.galbrun@aalto.fi

P. Miettinen is with Max Planck Institute for Informatics, Germany. email: pauli.miettinen@mpi-inf.mpg.de

This article is based on our recent tutorials called *An Introduction to Redescription Mining* at ECMLPKDD '16 and SDM '17, and on our upcoming book *Redescription Mining* [1].

[1] In ecology, the task is known as *bioclimatic niche (or envelope) finding* [2, 3].

[2] Slides available at http://siren.mpi-inf.mpg.de/tutorial/

In the table-based model, the data are arranged as a table (or tables; we will discuss that below). The rows of the table $\boldsymbol{D}$ correspond to the *entities* in the data and the columns correspond to the *attributes*; for example, in the bioclimatic niche finding example, the geographical regions are the entities, the table contains one row for each location where observations have been recorded, and the species and bioclimatic variables (that is, the observations) are the attributes. The value of attribute j in entity i is denoted as d_{ij}. The attributes can be of different types, such as binary, categorical, or numerical, and some entity–attribute values might be missing. In our example, the presence or absence of a species in a region constitutes a binary attribute, whereas the bioclimatic variables, such as temperatures or precipitations, are recorded as continuous numerical attributes.

A redescription is a pair of descriptions, and we formalize the descriptions as Boolean queries over the attributes. Each *predicate* in the queries assigns a truth value to (observed) entity–attribute pairs, that is, to the elements of a column of the data table. The queries over the predicates and their negations – together known as *literals* – in turn assign a truth value to each entity. The query can, in principle, be an arbitrary Boolean function of the literals, but it is common to restrict the queries to some *query language* for the sake of interpretability and efficiency of computation. Common query languages include *monotone conjunctive queries*, *linearly parsable queries* (where each variable can appear at most once and both conjunction and disjunction operators have the same precedence), and *tree-shaped queries* (a special case of disjunctive normal forms, encoding the paths from the root to the leaves in a decision tree).

Applying this formalism to our example niche redescription, the query corresponding to *'The areas inhabited by either Eurasian lynx or Canada lynx'* could be written as

$$\mathit{Eurasian\ lynx} \lor \mathit{Canada\ lynx}\ ,$$

and the query *'maximum March temperature ranges from* $-24.4\,°\mathrm{C}$ *to* $3.4\,°\mathrm{C}$*'* as

$$[-24.4 \le t_3^+ \le 3.4]\ .$$

To avoid tautological redescriptions (e.g. 'Eurasian lynx lives where Eurasian lynx lives'), we require that the queries do not share any attributes. In many applications, the attributes have a natural division into two disjoint sets. In our running example, the species form one set of attributes and the bioclimatic variables form the other set. In these cases, it is natural to model the data, not as a one, but as two data tables; one table for the species and one table for the bioclimatic variables, in our example. In this setup, the queries of a redescription are required to be over attributes from different tables.

The *support* of a query q, $\operatorname{supp}(q)$, is the set of entities (rows) that satisfy the query.[3] The support of the query *Eurasian lynx* $\lor$ *Canada lynx* contains the regions depicted in light red and in purple in Figure 1, while the support of the query $[-24.4 \le t_3^+ \le 3.4]$ contains the regions depicted in dark blue and in purple.

To form a good redescription, the queries should explain roughly the same entities, that is, their supports should be similar. The most common choice for measuring the similarity of the supports is the *Jaccard (similarity) index* J, defined as

$$J(p,q) = J(\operatorname{supp}(p), \operatorname{supp}(q)) = \frac{|\operatorname{supp}(p) \cap \operatorname{supp}(q)|}{|\operatorname{supp}(p) \cup \operatorname{supp}(q)|}\ .$$

The Jaccard index is by no means the only possible similarity measure, but it is by far the most common one. Its use can be motivated in many ways. For example, when using algorithms based on decision-tree (see Section III), it has a natural connection to the information gain splitting criteria [4]. On the other hand, if we consider redescription mining as mining bi-directional association rules (see again Section III), the Jaccard index of a redescription can be interpreted as the lower bound on the confidence of the corresponding association rules $\operatorname{conf}(p \Rightarrow q)$ and $\operatorname{conf}(q \Rightarrow p)$.

How similar should their supports be for the pair (p,q) to be considered a valid redescription is something the user must decide, depending on the data and her needs. Therefore, we say that the supports of p and q are similar enough if $J(p,q) \ge \tau$ for some user-specified constant $\tau \in [0,1]$, and write $p \sim q$.

We can now define what a redescription is. For data that consist of two tables, $\mathbf{D}_1$ and $\mathbf{D}_2$, a redescription is a pair of queries (p,q) expressed over attributes from $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively, such that $p \sim q$. In addition, a redescription might have to satisfy other constraints specified by the user, such as limitations on the size of the support, the maximum p-value, or the complexity of the queries (in terms of the number of variables involved, for instance). Then, the goal of *redescription mining* is to find all valid redescriptions $p_i \sim q_i$ that also satisfy the other potential constraints.

III. Algorithms

Readers familiar with classification and association rule mining might have noticed similarities between redescription mining and these two core data mining tasks. These two tasks provide basic techniques that have been adapted to develop algorithms for mining redescriptions.

Consider a case where one query is fixed and the goal is to find a matching query to make a good redescription: this can be seen as a binary classification problem [5]. This fact has inspired a family of iterative algorithms that alternate between the views to build the redescriptions. These algorithms derive target labels from a query obtained at a previous iteration and use classification techniques, typically decision tree induction, to build a matching query in the next iteration. The first algorithm proposed for redescription mining, the `CARTwheels` algorithm [4], is based on the idea of alternatively growing decision trees over one data table with only binary attributes. The decision-tree-based methods for arbitrary data types introduced by Zinchenko et al. [6] also belong to this family of redescription mining algorithms. Predictive clustering trees were used in a similar manner for mining redescriptions by Mihelčić et al. [7].

[3] Some sources call this set the *support set* and reserve the term support for what we call the *size* of the support.

On the other hand, association rule mining [8] can be seen as a precursor of redescription mining, with the latter allowing for more complex descriptions and focusing on equivalences instead of implications [4]. This inspired algorithms that first mine queries separately from the different views before combining the obtained queries across the views into redescriptions. The method proposed by Zaki and Hsiao [9] and the MID algorithm of Gallo et al. [10] both belong to this second family of algorithms. Along similar lines, Zaki and Ramakrishnan [11] studied exact and conditional redescriptions over binary data, focusing on conjunctive queries, while Parida and Ramakrishnan [12] studied the theory of exact redescriptions over binary attributes, where the queries are pure conjunctions, whether in monotone conjunctive normal form or monotone disjunctive normal form.

A third approach for mining redescriptions consists in growing them greedily. Such a strategy of progressively extending the descriptions by appending new literals to either query, always trying to improve the quality of the redescription, was first introduced as the Greedy algorithm of Gallo et al. [10]. Building upon this work, the ReReMi algorithm Galbrun and Miettinen [13] extended the approach to handle categorical and numerical attributes along with binary ones and use a beam search to keep the current top candidates at each step instead of focusing on the single best improvement.

The proposed algorithms can also be divided between exhaustive and heuristic strategies. Mine-and-pair algorithms based on association rule mining techniques are typically exhaustive. Alternating algorithms based on decision tree induction and algorithms that grow the queries greedily typically rely on heuristics.

While the first algorithms only considered binary attributes, more recent ones also allow to handle numerical and categorical attributes, possibly including missing entries. In this latter case, when calculating the supports of the queries and the similarity of the supports, a choice needs to be made about how to handle the entities for which the status of the queries cannot be determined due to missing values. Potential approaches include – but are not limited to – assuming that the queries always evaluate false on such entities [7] or assuming that they evaluate true or false depending on what is the most or the least favorable in terms of support similarity [13]. In fact, evaluating whether there is a way the query can evaluate true is NP-hard in general, though this is not the case with any of the query languages that are used with the existing algorithms. Of course, the actual mining algorithm also has to support missing values. For example, in algorithms using decision tree induction, the induction procedure must be able to handle missing values.

IV. Sets of redescriptions

Redescription mining, as defined above, is an exhaustive enumeration task, the goal being to output *all* valid redescriptions that satisfy the constraints. This is a common approach in data mining (cf. frequent pattern mining), but it can yield many redundant redescriptions. Filtering away the redundant redescriptions, however, requires us to define what redescriptions are redundant.

Perhaps the simplest approach is to consider the supports of the queries. We can order all (valid) redescriptions descending in their similarity, take the topmost redescription, move it to the list of non-redundant redescriptions, and mark the entities in its support 'used'. We can then re-evaluate the remaining redescriptions while only taking into account the non-used entities. All redescriptions that are deemed invalid (e.g. their support becomes too small or their Jaccard index too low) are considered redundant and removed. We repeat the process with the remaining redescriptions and entities until either the list of redescriptions or the set of entities becomes empty.

This simple approach can filter out too many redescriptions, as it only considers their support and not the attributes that appear in the descriptions. Kalofolias et al. [14] presented another approach for defining redundant redescriptions based on maximum-entropy distributions and the subjective interestingness approach of De Bie [15]. They model the data using a maximum-entropy distribution that is constrained so that the already-observed redescriptions have a high probability (or are certain) to occur. The other redescriptions are then ranked based on their likelihood of being true in a data following this model. The redescription that is the least-likely (i.e. the most surprising) is added as a constraint, the model is re-learned, and the remaining redescriptions are re-evaluated.

V. Tools

The Siren tool was developed for mining, visualizing, and interacting with redescriptions [16–18]. It provides a complete environment for redescription mining, from loading the data to finally exporting the results into various formats, through mining, visualizing, and editing the redescriptions.

Having good visualizations is crucial, of course, when designing a tool for visual data analysis. Indeed, visualization is the key to understanding the results of the mining process and we designed several visualizations for redescriptions. *Maps*, like the one presented in Figure 1, are a great way to understand where the queries hold (and do not hold), but require, naturally, that the entities are associated with geographical locations. *Parallel coordinates* plots are especially useful to understand the conditions appearing in the queries, as they allow to visualize the range of values selected by the predicates. Our example redescription depicted in a parallel coordinates plot is shown in Figure 2.

For redescriptions using decision tree induction and for tree-shaped queries more generally, *tree diagrams* reveal the tree structure underlying the queries, facilitating the interpretation of descriptions that can otherwise appear rather convoluted. A tree-shaped equivalent of our example redescription depicted in a tree diagram is shown in Figure 3.

Visualizations in Siren are linked, so that the user can highlight an entity across different visualizations of the same redescription, or interactively adjust the thresholds in the queries through the parallel coordinate plot, for instance. In addition, the tool allows to use different levels of automation when mining redescriptions, from letting the algorithm run fully automatically given a set of parameters, to letting the user edit the results fully manually, through partial automation where

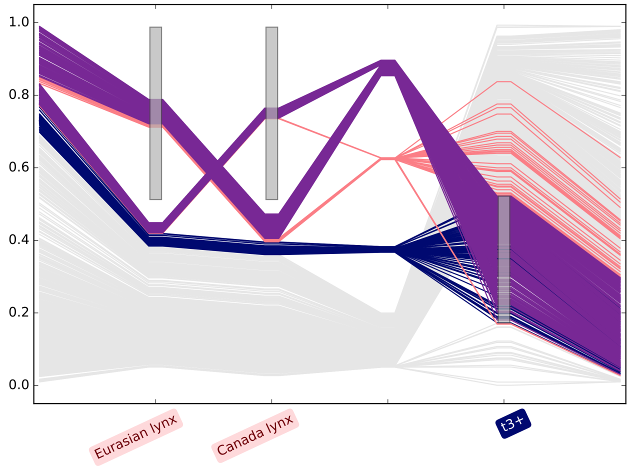

Figure 2. Parallel coordinates plot of our example redescription. Every line corresponds to one geographical location (entity) and the colours of the lines are as in Figure 1, except that grey correspond to locations where neither of the queries hold. The plot has three vertical axes corresponding to the three attributes in the redescriptions. The grey boxes in these axes correspond to the range of the values of the corresponding variable in the query; if a line passes through a gray box, the predicate corresponding to the attribute evaluates true for this entity.

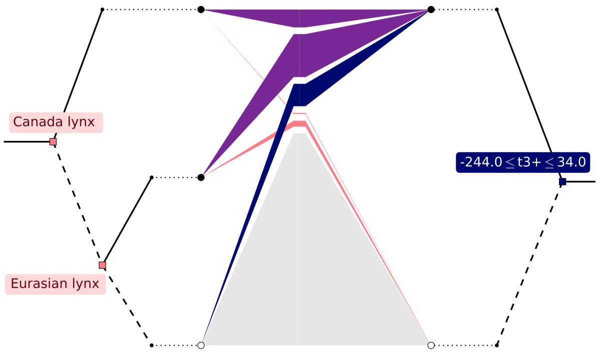

Figure 3. Tree diagram of a tree-shaped equivalent of our example redescription. Solid leaf nodes correspond to paths in the tree where the queries evaluate true, while empty leaves correspond to paths where the queries evaluate false. Lines between the two trees are as in Figure 2.

the algorithm extends and optimizes candidate redescriptions provided by the user.

Siren also allows to perfom support-based filtering on a set of redescriptions as explained in Section IV: the redescriptions are reranked and the redundant ones are marked.

Recently, Mihelčić and Šmuc [19] proposed a tool called InterSet for visualizing and working with sets of redescriptions. The tool allows to cluster redescriptions based on their shared attributes and shared entities. The user can also visualize the statistics of a set of redescriptions, such as the distribution of their Jaccard indices or of their pairwise support overlap, and filter the redescriptions based on those statistics.

VI. Applications

Redescription mining has been applied in various domains. Here, we present three examples from ecology, from biology and from social and policital sciences, respectively.

Instead of modelling the distributions of species directly, as in the niche finding example presented earlier, one might look at the distributions of functional traits of species. Galbrun et al. [20] consider dental traits of large plant eating mammals and bioclimatic variables (derived from temperature and precipitation records) from around the globe, looking for associations between teeth features and climate. The teeth of plant-eating mammals constitute an interface between the animal and the plant food available in its environment. Hence, teeth are expected to match the types of plant food present in the environment, and dental traits are thus expected to carry a signal of environmental conditions. In this study, three global zones are identified, namely a boreal-temperate moist zone, a tropical moist zone, and a tropical-subtropical dry zone, each associated to particular teeth characteristics and a specific climate.

Mihelčić et al. [21] use redescription mining to relate clinical and biological characteristics of cognitively impaired patients, with the aim of improving the early diagnosis of Alzheimer's disease. In this study, one data table consists of biological attributes derived from neuroimaging, from blood tests, and from genetic markers, for instance, while the other data table contains clinical attributes that record patients' answers to several questionnaires, observations by physicians, and results of cognition tests. The results obtained largely confirmed the findings of previous studies. In addition, they highlighted some additional biological factors whose relationships with the disease require further investigation, such as the pregnancy-associated plasma protein-A (PAPP-A), which they found to be highly associated with cognitive impairment in Alzheimer's disease.

Galbrun and Miettinen [22] applied redescription mining to analysing political opinion polls. Specifically, they used data from Finnish on-line voting advice applications, where candidates in the Finnish parliamentary elections have answered to a number of questions regarding their opinions on political matters, and had also provided socio-economical background data. Galbrun and Miettinen [22] analysed, first, the correlations between the socio-economical status and the political opinions of candidates, and, second, compared the answers of candidates who run for both 2011 and 2015 elections between these years. Their findings partially followed the party platforms, but they also found unsuspected connections; for example, candidates who were over 37 years old or who had children were not strongly supporting legalizing euthanasia, and vice versa.

VII. Related methods

As we have seen, the work on redescription mining has significantly expanded and diversified since the task was first formalized by Ramakrishnan et al. [4]. Problem variants have also been introduced: *storytelling* aims at building a chain of redescriptions linking given objects or queries while *relational* *redescription mining* aims to find redescriptions in heterogenous networks.

Beside classification and association rule mining (see Section III), the task also has connections with subgroup discovery, clustering and multi-view approaches, in particular.

In subgroup discovery [23], the input contains features and a target variable over observations, and the goal is to find queries that describe groups that have an ‘interesting’ behaviour in the target variable, that is, groups of entities that have different statistical properties (e.g. average) in the target variable when compared to the rest of the observations.

Clustering is a classical unsupervised data analysis method with the goal of grouping the entities in such a way that entities in the same group are as similar to each other as possible, and the objects in different groups are as dissimilar from each other as possible. A query can be interpreted as selecting a subset of the attributes and a group of entities that are in some sense ‘similar’ to each other, although not in the classical sense (e.g. of having short Euclidean distance). Among clustering techniques, redescription mining is most related to subspace clustering [24] and biclustering [25].

An important feature of redescriptions is their ability to describe data from different points of view, i.e. their ‘multi-view’ aspect. Other examples of *multi-view data mining methods* include *multi-view clustering* [26], where the attributes are divided into two views and the clustering is done separately over each view; *multi-view subgroup discovery* [27], where the subgroup discovery is done over multiple views; and various *multi-view matrix and tensor factorization* [28–30], which use (partially) the same factors to decompose multiple matrices or tensors.

VIII. Conclusion and future work

Redescription mining is a powerful data analysis technique that is gathering wider interest, among data analysis researchers and practitioners alike. The availability of efficient algorithms that can handle heterogeneous data types has undoubtably contributed to the increasing adoption. Yet, redescription mining is, in many ways, in its infancy, and there are still many interesting open questions to be addressed. Developing redescription mining methods that work over time series data is one important future direction. Another interesting direction is to add predicates that are functions of the attributes, such as square roots, logarithms, squares, and so on, and perhaps also multivariate composite attributes. This would naturally allow the query to capture more complex structures, but the exact functions would have to be application-dependant. Finally, redescription mining could also be extended to more complex data (*relational redescription mining* [31] can be seen as one step in that direction), such as graphs and multimodal (e.g. tensor) data.

References

[1] E. Galbrun and P. Miettinen, *Redescription Mining*. Springer, 2018.

[2] J. Soberón and M. Nakamura, “Niches and distributional areas: Concepts, methods, and assumptions,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. Supplement 2, pp. 19 644–19 650, 2009.

[3] J. Grinnell, “The niche-relationships of the california thrasher,” *The Auk*, vol. 34, no. 4, pp. 427–433, 1917.

[4] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R. F. Helm, “Turning CARTwheels: An alternating algorithm for mining redescriptions,” in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’04)*, 2004, pp. 266–275.

[5] C. C. Aggarwal, “Chapter 10,” in *Data Mining: The Textbook*. Cham: Springer, 2015.

[6] T. Zinchenko, E. Galbrun, and P. Miettinen, “Mining predictive redescriptions with trees,” in *IEEE International Conference on Data Mining Workshops*, 2015, pp. 1672–1675.

[7] M. Mihelčić, S. Džeroski, N. Lavrač, and T. Šmuc, “Redescription mining with multi-target predictive clustering trees,” in *Proceedings of the 4th International Workshop on the New Frontiers in Mining Complex Patterns (NFMCP’15)*, 2016, pp. 125–143.

[8] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD’93)*, 1993, pp. 207–216.

[9] M. J. Zaki and C.-J. Hsiao, “Efficient algorithms for mining closed itemsets and their lattice structure,” *IEEE Trans. Knowl. Data En.*, vol. 17, no. 4, pp. 462–478, 2005.

[10] A. Gallo, P. Miettinen, and H. Mannila, “Finding subgroups having several descriptions: Algorithms for redescription mining,” in *Proceedings of the 8th SIAM International Conference on Data Mining (SDM’08)*, 2008, pp. 334–345.

[11] M. J. Zaki and N. Ramakrishnan, “Reasoning about sets using redescription mining,” in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’05)*, 2005, pp. 364–373.

[12] L. Parida and N. Ramakrishnan, “Redescription mining: Structure theory and algorithms,” in *Proceedings of the 20th National Conference on Artificial Intelligence and the 7th Innovative Applications of Artificial Intelligence Conference (AAAI’05)*, 2005, pp. 837–844.

[13] E. Galbrun and P. Miettinen, “From black and white to full color: Extending redescription mining outside the Boolean world,” *Stat. Anal. Data Min.*, vol. 5, no. 4, pp. 284–303, 2012.

[14] J. Kalofolias, E. Galbrun, and P. Miettinen, “From sets of good redescriptions to good sets of redescriptions,” in *Proceedings of the 16th IEEE International Conference on Data Mining (ICDM’16)*, 2016, pp. 211–220.

[15] T. De Bie, “Maximum entropy models and subjective interestingness: an application to tiles in binary databases,” *Data Min. Knowl. Discov.*, vol. 23, no. 3, pp. 407–446, 2011.

[16] E. Galbrun and P. Miettinen, “A case of visual and interactive data analysis: Geospatial redescription mining,” in *Proceedings of the ECML PKDD 2012 Workshop on Instant and Interactive Data Mining (IID’12)*, 2012, Accessed 25 Oct 2017. [Online]. Available: http://adrem.ua.ac.be/iid2012/papers/galbrun_miettinen-visual_and_interactive_geospatial_redescription_mining.pdf

[17] ——, “Siren: An interactive tool for mining and visualizing geospatial redescriptions [demo],” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’12)*, 2012, pp. 1544–1547.

[18] ——, “Interactive redescription mining,” in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD’14)*, 2014, pp. 1079–1082.

[19] M. Mihelčić and T. Šmuc, “InterSet: Interactive redescription set exploration,” in *Proceedings of the 19th International Conference on Discovery Science (DS’16)*, vol. 9956, 2016, pp. 35–50.

[20] E. Galbrun, H. Tang, M. Fortelius, and I. Žliobaitė, “Computational biomes: The ecometrics of large mammal teeth,” *Palaeontol. Electron.*, 2017, submitted.

[21] M. Mihelčić, G. Šimić, M. Babić-Leko, N. Lavrač, S. Džeroski, and T. Šmuc, “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, 2017.

[22] E. Galbrun and P. Miettinen, "Analysing political opinions using redescription mining," in *IEEE International Conference on Data Mining Workshops*, 2016, pp. 422–427.

[23] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," in *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'97)*, vol. 1263, 1997, pp. 78–87.

[24] P. Kröger and A. Zimek, "Subspace clustering techniques," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. New York: Springer, 2009, pp. 2873–2875.

[25] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE Trans. Comput. Bio. Bioinform.*, vol. 1, no. 1, pp. 24–45, 2004.

[26] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04)*, 2004, pp. 19–26.

[27] L. Umek, B. Zupan, M. Toplak, A. Morin, J.-H. Chauchat, G. Makovec, and D. Smrke, "Subgroup discovery in data sets with multi-dimensional responses: A method and a case study in traumatology," in *Proceedings of the 12th Conference on Artificial Intelligence in Medicine (AIME'09)*, vol. 5651, 2009, pp. 265–274.

[28] P. Miettinen, "On finding joint subspace Boolean matrix factorizations," in *SIAM International Conference on Data Mining (SDM'12)*, 2012, pp. 954–965.

[29] S. K. Gupta, D. Phung, B. Adams, and S. Venkatesh, "Regularized nonnegative shared subspace learning," *Data Min. Knowl. Disc.*, vol. 26, no. 1, pp. 57–97, 2013.

[30] S. A. Khan and S. Kaski, "Bayesian multi-view tensor factorization," in *Proceedings of the 2014 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'14)*, 2014, pp. 656–671.

[31] E. Galbrun and A. Kimmig, "Finding relational redescriptions," *Mach. Learn.*, vol. 96, no. 3, pp. 225–248, 2014.