Interactive Redescription Mining

Esther Galbrun*
Department of Computer Science
Boston University
Boston, MA, USA
galbrun@cs.bu.edu

Pauli Miettinen
Max Planck Institute for Informatics
Saarbrücken, Germany
pmiettin@mpi-inf.mpg.de

ABSTRACT

Exploratory data analysis consists of multiple iterated steps: a data mining method is run on the data, the results are interpreted, new insights are formed, and the resulting knowledge is utilized when executing the method in a next round, and so on until satisfactory results are obtained.

We focus on redescription mining, a powerful data analysis method that aims at finding alternative descriptions of the same entities, for example, ways to characterize geographical regions in terms of both the fauna that inhabits them and their bioclimatic conditions, so-called *bioclimatic niches*.

We present Siren, a tool for interactive redescription mining. It is designed to facilitate the exploratory analysis of data by providing a seamless environment for mining, visualizing and editing redescriptions in an interactive fashion, supporting the analysis process in all its stages. We demonstrate its use for exploratory data mining.

Simultaneously, Siren exemplifies the power of the various visualizations and means of interaction integrated into it; Techniques that reach beyond the task of redescription mining considered here, to other analysis methods.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data Mining*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Data Visualization Methods*

Keywords

Redescription Mining; Interactive Data Mining; Visualization; Brush-and-Link; Parallel Coordinates

1. INTRODUCTION

The goal of data mining is to find surprising but useful new patterns and relations from the data. This process involves multiple steps such as preparing the data, selecting and applying the data mining methods, and interpreting the results. Furthermore, it is generally iterative, with the results of the first round used to inform the decisions made in the second round, *et cetera*.

The analysis is — or at least should be — done by a domain expert, as only a domain expert can judge what kind of results are surprising or useful. But to allow the domain expert to perform the data analysis as effectively as possible, the process should be as seamless as possible. Typically, it is unreasonable to assume that the domain experts are proficient on using half a dozen different programs, often with arcane command-line interfaces, to deploy the full analysis process. There are two common ways to achieve the desired consistency: either via a workflow integrated into a well-known general-purpose analysis framework, such as R or Matlab, or via a special-purpose tool specifically designed for the task at hand.

Siren takes the latter approach: it is a tool for interactively mining and visualizing redescriptions. *Redescription mining* [9] is a data analysis method that aims to find different ways of characterizing the same things and, vice versa, to find things that admit the same alternative characterizations. For instance, consider the task of bioclimatic niche finding [10]: we are given a set of geographical regions (e.g. 50 km squares) with information about the species inhabiting these regions, on one hand, and the bioclimatic conditions (e.g. temperature and precipitation) encountered in each of them, on the other hand. The task is then to find a set of species and a set of bioclimatic conditions such that the bioclimatic conditions explain exactly (or as well as possible) the areas where the species live, and vice versa. Such a pair forms a redescription since it characterizes (approximately) the same area in two different ways.

An example redescription from bioclimatic niche finding characterizes Scandinavia and the Baltic as the habitat of the Moose, on one hand, and in term of its specific cold climate, on the other hand. A map plot of this redescription can be seen in the foreground of Figure 1. We will use niche finding as our example application throughout this paper, although we emphasize that redescription mining is not limited to any single application, or to any specific domain. Examples of other applications and redescription mining in general are discussed further in [9, 1], for instance.

Siren can also be seen as a case study of the combination of various visualizations and interaction techniques into a powerful data analysis tool. Such integration could potentially also benefit methods other than redescription mining.

*The work was done while the author was a doctoral student at the University of Helsinki, with support from the Academy of Finland, Grant 255675.

SIGMOD'14, June 22–27, 2014, Snowbird, UT, USA.

ACM 978-1-4503-2376-5/14/06 ...$15.00.
http://dx.doi.org/10.1145/2588555.2594520.

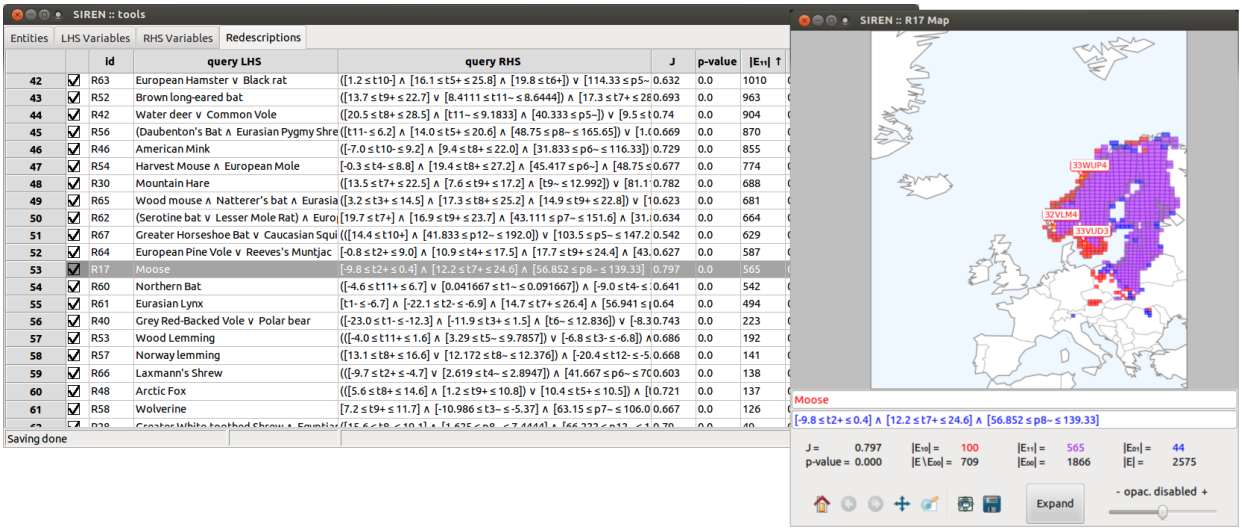


Figure 1: The Siren interactive mining and visualization tool. The panel in the background contains a list of redescriptions while the foreground panel displays the map plot of a selected redescription.

In the next section we provide a very brief introduction to redescription mining and cover some related work. Section 3 explains the main features of SIREN and why we think they are important. Section 4 outlines techniques used for implementing this tool, and Section 5 is a brief walk-through of the demonstration.

We presented a preliminary version of SIREN in [3]. That version introduced the GUI and preliminary connections to the mining algorithm. However, it lacked visualizations other than plotting results on a map and did not have the more sophisticated interactive features of the current version. Furthermore, the underlying architecture has been significantly improved. In this paper we focus on the new features, except when necessary to understand the broader functionality. In other words, features discussed here are new compared to [3], unless otherwise mentioned.

More details about SIREN, including videos, a user guide and further references on redescription mining are available on the associated webpage.[1]

2. REDESCRIPTION MINING

Redescription mining is a descriptive data analysis task. It aims at simultaneously finding multiple descriptions of a subset of entities which is not previously specified [9]. This is in contrast with other methods like *emerging pattern mining*, *contrast set mining*, or *subgroup discovery* (see [7] for a unifying survey) or general classification methods, where target subsets of entities are specified via labels.

We consider data that contains entities with two sets of characterizing variables, such as the fauna and bioclimatic conditions. In this setting, the redescription mining task consists in finding a pair of queries, one query for each set of variables, such that both queries describe (almost) the same set of entities. We refer to the two sets of variables as left and right hand side data, and the queries over them, respectively, as left and right hand side queries. A redescription is simply a pair of queries over variables from the two sets. The support of a query is the subset of entities for which the query holds true. The *accuracy* of a redescription is measured by the *Jaccard coefficient* (denoted as J) of the supports of its two queries; p-values indicating how likely it is to observe such an overlap for independent queries can be used to reject uninteresting redescriptions.

Several algorithms for redescription mining have been proposed over the years and in principle SIREN can use any of them as the method to mine the redescriptions. To the best of our knowledge, however, only the REREMI algorithm [2] can handle numerical or categorical data and missing values. It is therefore our algorithm of choice.

Our example application here is *bioclimatic niche-finding*. The aim is to determine the bioclimatic constraints that must be met for a certain species to survive, called the species' bioclimatic envelope, or niche [4]. Finding such envelopes could help, for instance, to predict the results of global warming [8]. Unlike other tools developed over the past ten years to model the bioclimatic envelope, such as BIOMOD [11], redescription mining allows automatically finding both the set of species and their envelope.

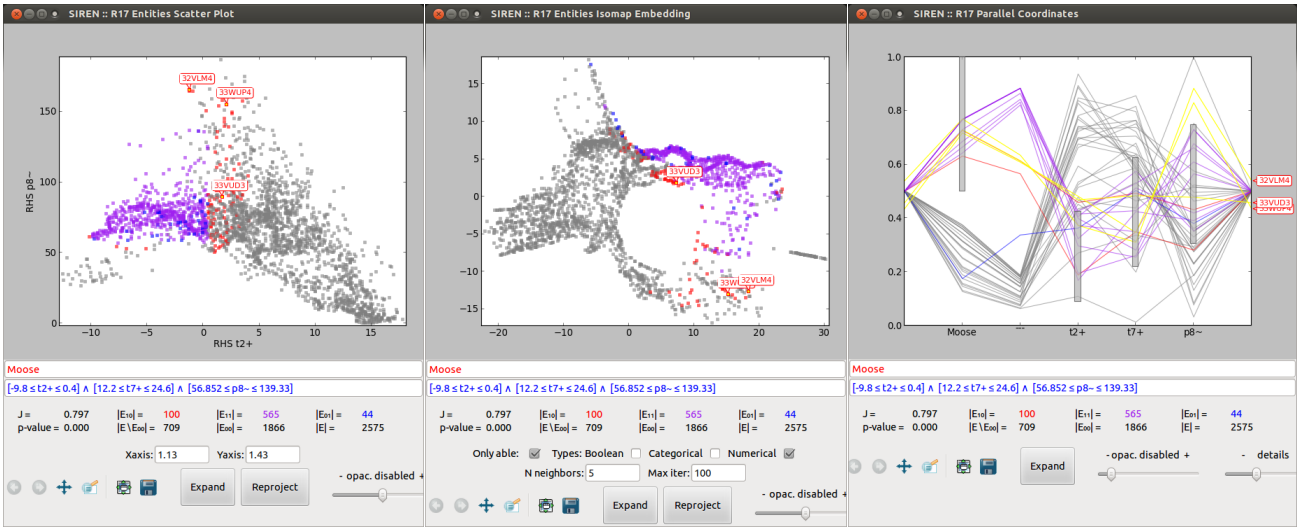
While we cannot cover other applications in this demonstration proposal due to space constraints, we refer the interested user to the aforementioned web page where we give three other example applications from different domains.

3. THE FEATURES OF SIREN

SIREN 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. SIREN allows for a seamless interaction with both the mining and visualization, enabling the user to interactively edit the redescriptions in the visualizations and to call the mining algorithm, e.g. to extend the current results. In what follows, we cover some of the main features of SIREN and argue why they are important.

Mining. Obviously, the core of SIREN is mining the redescriptions. Having just a GUI atop the mining algorithm is not enough, however, in particular because mining all the redescriptions from a dataset can be a time-consuming task.

[1] http://www.cs.helsinki.fi/u/galbrun/redescriptors/siren/



(a) (b) (c)

Figure 2: Alternative visualizations of the redescription shown in the foreground panel of Figure 1. Three selected entities are highlighted across the different visualizations.

To avoid extensive waiting times, mining redescriptions in Siren is an asynchronous any-time process, meaning that the user will start seeing results from the mining algorithm as soon as the first (partial) results are ready. The user can start working with and editing these redescriptions while the mining algorithm continues in the background. While Siren was capable of doing this already in [3], we have improved the implementation of the asynchronous computations to take full benefit from the multiple cores in modern CPUs. Furthermore, the new asynchronous model allows Siren to off-load heavy computational tasks to an external server.

When exploring the data, the user often wants to extend existing redescriptions, either those returned by the algorithm or else some queries she has constructed herself, e.g. to test a hypothesis about the data. Siren allows to use an existing redescription as a starting point for the mining process. The user can also select only one side of the redescription to be extended. In bioclimatic niche finding, for example, this can be used to see how good a bioclimatic envelope the algorithm can find for a specific combination of species.

Furthermore, the user may specify a subset of entities that she wants to be emphasized during the mining process. For example, if the user wants to add some entities (in this case, geographic areas) to the support of the redescription shown in Figure 1, she can ask Siren to extend the given redescription with emphasis on including them in the support. In other words, the user can manipulate a redescription both through its queries and through its support.

The user can also disable variables and/or entities so that they will not be used in the mining process. For example, there can be known anomalous areas (e.g. coastal regions or valleys in mountain ranges) that the user might want to exclude, to prevent them from affecting the algorithm. The user can also disable the entities after the mining is done, in which case Siren will automatically update the support and accuracy of all redescriptions.

Visualizing. Visualization is the key to understanding the results of the mining process. As, for example, Heer and Shneiderman [5] argued, having multiple concurrent visualizations of the same data reinforces their explanatory power. While the preliminary version of Siren was only capable of plotting redescriptions on a map, multiple different visualizations of the entities are now available. They are also applicable to data without spatial coordinates, thereby lifting the restriction to geospatial data.

The simplest of these visualizations are the various projections of the data into the 2D space. Different types of projections have been studied intensively, and Siren provides a number of them, including Karhunen–Loève transform (i.e. PCA), multi-dimensional scaling and various scatter plots. Figure 2 (a) and (b) are two alternative projections for the same example redescription, namely a scatter plot and an isomap embedding. In all visualizations, colors encode whether an entity belongs to the support of the left hand side query (denoted as E_{10}), the right hand side query (E_{01}), or both (E_{11}). Here, we use red, blue, and purple respectively. Adding new projections is straightforward due to the modular nature of Siren. Computing these projections for larger datasets can be computationally intensive, but that too can be off-loaded to an external server.

Parallel coordinates plots [6] are particularly suited for visualizing redescriptions. In such a plot, the entities are represented by lines going through a series of parallel vertical axes, one for each literal appearing in the queries. The position where a line crosses an axis indicates the value of the associated variable for the corresponding entity. For each literal, the range of values that make the truth assigment hold is represented by a grey interval box. An extra axis separates the two sides and registers the support of the queries. Figure 2 (c) shows an example of a parallel coordinates plot. The same color code is used for the entities as with other visualizations. In addition, grey lines represent entities that do not support either query (E_{00}).

An example use case for the multiple visualizations is to study why some entities are not included in the support of both queries. To facilitate the analysis of such situations,

SIREN provides brush-and-link capabilities. The user can select (brush) an entity or a set of entities by clicking them or by drawing a polygon around them on any of the visualizations and the corresponding entities are highlighted in all associated visualizations. This way, the user can use the map to select areas of interest then use the 2D projections to quickly see whether these areas behave anomalously compared to other areas. In our example redescription, some areas where the Moose lives do not appear in the support of the climate query (drawn in red). We select three of them and notice from the two projections that the coastal areas of Norway (32VLM4 and 33WUP4) have a climatic profile which is clearly distinct from most covered areas, unlike the area in southern Sweden (33VUD3). From the parallel coordinates plot we then determine that raising the upper bound on February maximum temperature would allow to cover the latter area.

Editing. A redescription can be edited from any visualization by using the text fields under the plot. Thus, the user might adjust the conditions in the queries, add or remove variables, as well as build entirely new redescriptions by hand. Upon editing a query, the plot and the statistics of the redescriptions are recomputed to account for the modifications. This behavior was already implemented in the preliminary version of SIREN.

Now, it is also possible to edit the queries directly from the parallel coordinates plot by dragging the interval boxes to modify the bounds or categories for the variables. Continuing with our example, we can simply drag the top of the second grey box up to the highlighted line to include the area in the support (and some other areas as well). This will trigger the update of all the associated visualizations and re-computation of the statistics of the redescription.

4. IMPLEMENTATION DETAILS

SIREN is available as a desktop program. Both SIREN and REREMI are implemented in Python to ensure cross-platform compatibility. The interface is built with the `wxPython` Open Source cross-platform GUI toolkit and relies on the `matplotlib` library for the visualizations. The 2D projections are implemented as modular functions that take a `numpy` matrix as an input and output a list of 2D points. Most of the projections call the corresponding routines from the `scikit-learn` package.

In SIREN, all mining and projection tasks are handled in separate processes, using Python's standard `multiprocessing` library. Multiple independent computations can thus be run simultaneously, fully exploiting modern multi-core CPU architectures. The tasks can either be run locally or be off-loaded to an external server. The communication between SIREN and the computational server are also handled with `multiprocessing` and we use a server-client architecture, so that a single server can serve multiple SIREN clients by spawning a new worker for each job request.

5. DEMONSTRATION SCENARIO

In the demonstration we will present SIREN following a usage scenario of finding bioclimatic niche with redescription mining, unless the audience prefers some other application.

After a brief explanation of the concepts underlying redescription mining and the chosen application scenario, we will demonstrate the features of SIREN. In particular, we will highlight how the multiple visualizations together with the brush-and-link workflow support a better understanding and interpretation of the redescriptions and how the interface intuitively allows to impart domain knowledge to the mining process, so as to improve the results in subsequent iterations.

For the interest of time, we will start with few pre-mined redescriptions, but will then invite the audience to edit these redescriptions and to use the mining algorithm to extend them. Since SIREN is a fully functional program that anyone can download from the web page, the demonstration can be fully interactive, with the users directly trying out the tool, possibly on their own datasets.

We have prepared a short video showing a sample walkthrough of the demonstration. It can be seen in the aforementioned accompanying web page.

6. CONCLUSIONS

SIREN provides a seamless environment for interactively mining, visualizing, and editing redescriptions. Its multiple visualizations, coordinated plots, brush-and-link workflow, and tight integration to the mining algorithm make it a very effective exploratory data analysis tool.

This tool also provides a case study on the power of various means of visualization and interaction which are potentially relevant in the context of other exploratory data analysis methods as well.

7. REFERENCES

[1] E. Galbrun. *Methods for Redescription Mining*. PhD thesis, University of Helsinki, Helsinki, Dec. 2013.

[2] E. Galbrun and P. Miettinen. From Black and White to Full Colour: Extending Redescription Mining Outside the Boolean World. *Stat. Anal. and Data Min.*, 5(4):284–303, 2012.

[3] E. Galbrun and P. Miettinen. Siren: An interactive tool for mining and visualizing geospatial redescriptions – Demo. In *KDD*, pages 1544–1547, 2012.

[4] J. Grinnell. The niche-relationships of the California Thrasher. *The Auk*, 34(4):427–433, 1917.

[5] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis. *Comm. ACM*, 55(4):45–54, Apr. 2012.

[6] A. Inselberg. *Parallel coordinates: Visual multidimensional geometry and its applications*. Springer, 2009.

[7] P. K. Novak, N. Lavrac, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *J. Mach. Learn. Res.*, 10:377–403, 2009.

[8] R. G. Pearson and T. P. Dawson. Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecol. Biogeogr.*, 12:361–371, 2003.

[9] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R. F. Helm. Turning CARTwheels: An alternating algorithm for mining redescriptions. In *KDD*, pages 266–275, 2004.

[10] J. Soberón and M. Nakamura. Niches and distributional areas: Concepts, methods, and assumptions. *PNAS*, 106(Supplement 2):19644, 2009.

[11] W. Thuiller, B. Lafourcade, R. Engler, and M. B. Araújo. Biomod – a platform for ensemble forecasting of species distributions. *Ecography*, 32(3):369–373, 2009.