Siren: An Interactive Tool for Mining and Visualizing Geospatial Redescriptions

[Demo]

Esther Galbrun
Helsinki Institute for Information Technology
Department of Computer Science
University of Helsinki, Finland
galbrun@cs.helsinki.fi

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

ABSTRACT

We present SIREN, an interactive tool for mining and visualizing geospatial redescriptions. Redescription mining is a powerful data analysis tool that aims at finding alternative descriptions of the same entities. For example, in biology, an important task is to identify the bioclimatic constraints that allow some species to survive, that is, to describe geographical regions in terms of both the fauna that inhabits them and their bioclimatic conditions.

Using SIREN, users can explore geospatial data of their interest by visualizing the redescriptions on a map, interactively edit, extend and filter them[1].

To demonstrate the use of the tool, we focus on climatic niche-finding over Europe, as an example task. Yet, SIREN is by no means limited to a particular dataset or application.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data Mining*

General Terms

Algorithms, Design

Keywords

Redescription Mining, Interactive Data Mining

1. INTRODUCTION

Finding multiple ways to characterize the same entities is a problem that appears in many areas of science. In medical sciences, one might want to find a subset of patients sharing similar symptoms and similar genes. Describing geographical regions in terms of both their bioclimatic conditions and the fauna that inhabits them is another example — and a task of great importance for biologists. A simple example of a redescription in this setting could state that areas where Moose live are areas where February's maximum temperature is between -10 and 0 degrees Celsius and July's maximum temperature between 12 and 25 degrees Celsius. This is the redescription shown in the front panel of Figure 1.

However, finding such alternative descriptions of the data generally requires to manually fix the query on one side before looking for the best matching query on the other side, using some type of classification. This typically prevents queries involving more than one variable to be tested. *Redescription Mining* aims at solving this tedious task by automatically identifying the best redescriptions.

More formally, we consider data that contains entities with two sets of characterizing variables, e.g. the fauna and the bioclimatic conditions. 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. The task consists in finding a pair of queries, one query for both sets of variables, such that both queries describe (almost) the same set of entities. Section 4 presents requisite background information about that problem.

The results of redescription mining, the redescriptions, can be approached from two points of view. On one hand, by considering the variables and conditions appearing in the queries, which provide valuable information in themselves; On the other hand, by studying the support set of the redescriptions, i.e. the subset of entities where both queries of a redescription hold. When the data is geospatial, that is, the entities are connected to geographical locations, the latter approach becomes even more important. A meaningful geospatial redescription should define coherent areas using expressive queries. The goal of SIREN is to facilitate the analysis of redescriptions using both of the approaches simultaneously.

Mining data is generally an iterative process, the results obtained at one step giving rise to hypotheses which will be tested at a further step, and redescription mining is no exception. Providing means to the user to easily interact with the mining process greatly improves the analysis. When dealing with geospatial data, visualizing the results on a map is crucial in order to interpret them. To answer these

[1]More details about SIREN's features, additional screenshots and a demonstration video are available online at `http://www.cs.helsinki.fi/u/galbrun/redescriptors/siren/`.

KDD'12, August 12–16, 2012, Beijing, China.

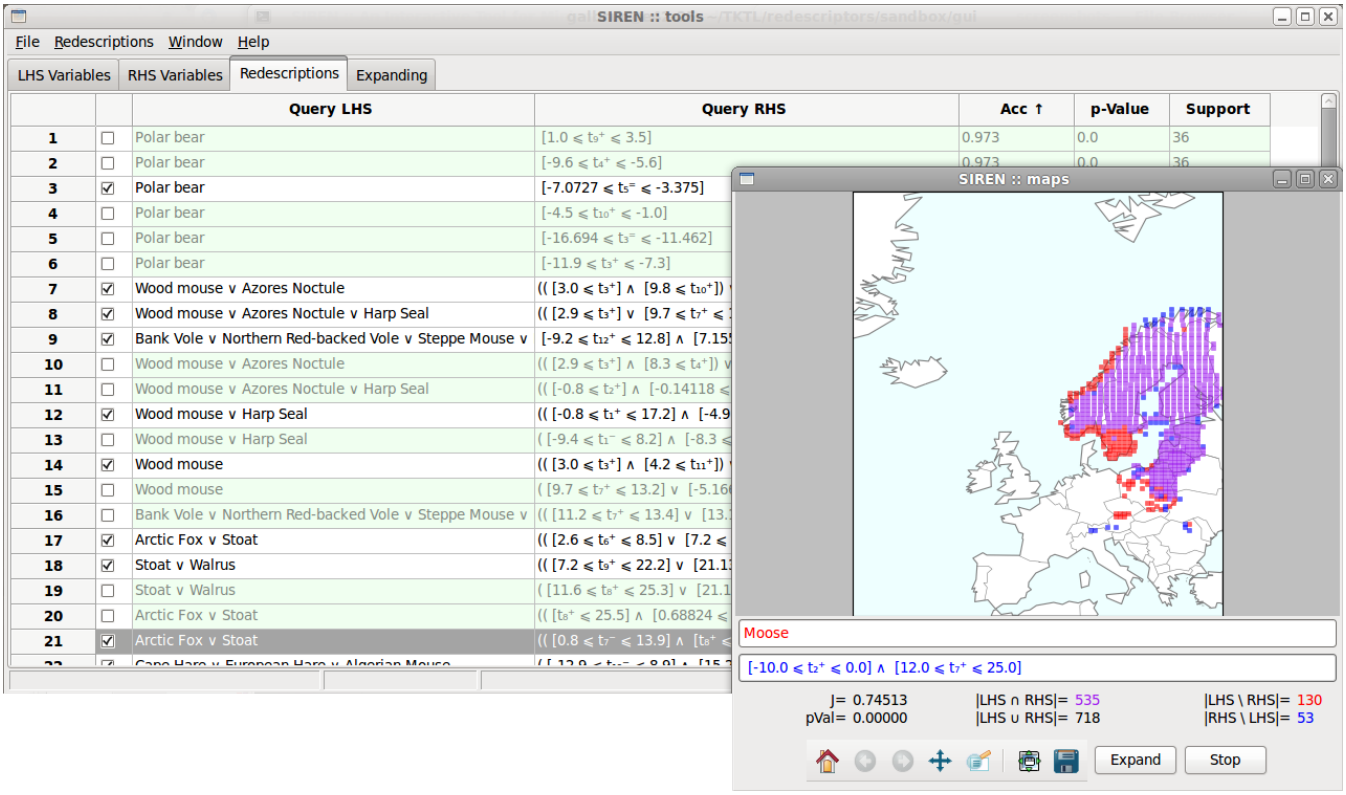

Figure 1: The Siren interactive mining and visualization tool. The panel in the background contains a list of redescriptions while the foreground panel displays the map of a selected redescription. In this example, left hand side queries are over fauna while right hand side queries are over monthly bioclimatic conditions, that is, temperatures and precipitation.

needs, we present SIREN, an interactive tool for mining and visualizing geospatial redescriptions.

Alternatively, experimenting with SIREN provides an efficient way for unfamiliar users to learn about redescription mining. The underlying concepts can be easily understood while visualizing the redescriptions and interacting with the system. Hence, it can also be used for educational purposes.

2. EXAMPLE APPLICATIONS

Our main example for demonstrating SIREN is an application to *bioclimatic niche-finding*.

Indeed, an important problem in biology, niche-finding, is a particular instance of redescription mining. The bioclimatic constraints that must be met for a certain species to survive constitute that species' bioclimatic envelope, or niche [3]. Finding such envelopes can help, e.g. to predict the results of global warming [9]. A number of methods, involving regression, neural networks, and genetic algorithms (see [11]) have been developed over the past ten year to model the bioclimatic envelope, BIOMOD [12] being a good example of a modelling tool used in this domain. But to the best of our knowledge, none of these methods allows automatically finding both the set of species and their envelope.

We give an example of the application of SIREN on this task using data that describes spatial areas of Europe, squares of side roughly 50 kilometers. The left hand side data contains information about the mammals living in these areas, while the right hand side consists of bioclimatic variables. The data comes from two publicly available datasets: European mammal atlas [6] and Worldclim climate data [4].

Nonetheless, SIREN is a flexible tool that can be used with different datasets from various application domains. For instance, it could help in sociological studies, with the exploration of statistical and political data. On the web page, we outline a usage scenario based on census statistics and electoral campaign funding of U.S. continental counties, as an alternative example.

3. USE-CASE SCENARIO

We exemplify the usage of SIREN by going through a generic work-flow of mining geospatial redescriptions, detailing typical steps in the process. A screenshot of the system, displaying a list of redescriptions with one particular redescription plotted on a map is shown in Figure 1.

Initial redescription mining.

A natural starting point for the analysis of any given data is to use a redescription mining algorithm to find an initial set of redescriptions. This can be done withing SIREN by running the extension mechanism on an empty redescription.

Extending a redescription.

Sometimes the user wants to focus only on one of the queries, on some particular variable of interest or on a part of an existing redescription. SIREN allows the user to auto-

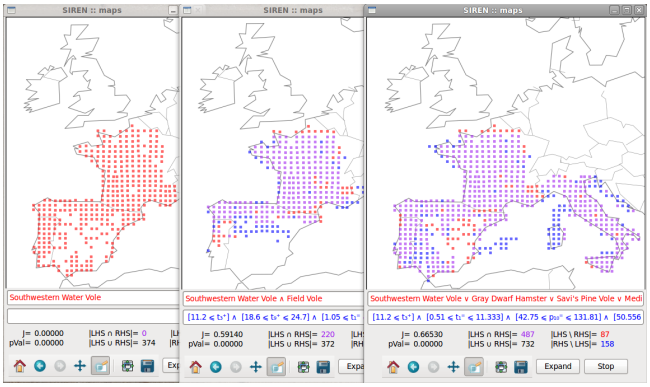

Figure 2: Several map panels. Comparing intermediate extensions automatically generated for a chosen starting variable. Red, blue and purple represents areas where only the left hand side query holds, only the right hand side query holds and where both queries hold, respectively.

matically extend a given redescription, i.e. let the algorithm add new literals to the queries to make the redescription as accurate as possible.

The extension mechanism of SIREN is based on the beam search implemented in the REREMI algorithm [1]. In this case, the intermediate redescriptions explored during the search are returned at each step, allowing to study more specific alternative extensions to a redescription that were discarded from the beam because they were not among the best extensions at some point of the search.

In the climatic niche-finding task, for instance, we might select a species, say, the Southwestern Water Vole and look for best extensions starting from that single variable. Returned extensions can be visualized side by side and compared as shown in Figure 2. Here, the best found extension has accuracy 0.665 (per Jaccard coefficient):

$$\text{Southwestern Water Vole} \vee \text{Gray Dwarf Hamster} \vee \text{Savi's Pine Vole} \vee \text{Mediterranean Monk Seal}$$

$$[11.2 \leq t_3^+] \wedge [0.51 \leq t_1^= \leq 11.333] \wedge [42.75 \leq p_{10}^= \leq 131.81] \wedge [50.556 \leq p_{11}^= \leq 176.75],$$

This redescription indicates that areas where any of the four species lives correspond to areas where the maximum temperature in March is above 11.2 degrees Celsius, the average temperature in January between 0.51 and 11.333 degrees Celsius and the average precipitations in October and November range from 42.75 to 131.81 millimeters and from 50.556 to 176.75 millimeters, respectively.

Editing a redescription.

It is typical that the user wants to edit some of the obtained redescriptions. For example, some results might be overly complex, or have exceedingly precise boundaries for numerical variables. The user can easily select a redescription to modify, open it in a map panel and edit it. Boundaries can be altered, literals added or removed. SIREN updates the map and important statistics (accuracy, p-value, etc.) of the redescription, allowing the user to see the effects of the modifications immediately and verify, e.g. whether the new redescription would still be acceptably accurate.

Continuing with our example above, we might want to reduce the precision of the climatic constraints to integers. We could edit the query as follows:

$$[11 \leq t_3^+] \wedge [0 \leq t_1^= \leq 12] \wedge [42 \leq p_{10}^= \leq 132] \wedge [50 \leq p_{11}^= \leq 177],$$

and obtain a redescription of slightly decreased accuracy.

Using subsets of variables.

SIREN allows the user to specify variables to temporarily avoid when extending or mining redescriptions.

For example, when two variables are highly correlated, several redescriptions might contain them both. The user, however, might want to consider redescriptions with only one of these variables, not both. SIREN makes that simple: the user only has to select a redescription, remove the unwanted variable from the redescription and unselect it from the list of variables, then extend the redescription again.

Alternatively, in our running example, we might want to force the algorithm to search alternative redescriptions that do not involve any precipitation. For that purpose, we simply unselect all such variables before running the extension anew. We will obtain the best extensions containing only temperatures in the bioclimatic query, such as the following redescription of accuracy 0.653:

$$\text{Southwestern Water Vole} \vee \text{Cape Hare} \vee \text{Savi's Pine Vole} \vee \text{Mediterranean Monk Seal}$$

$$([11.2 \leq t_3^+] \wedge [20.1 \leq t_7^+ \leq 32.9] \wedge [0.51 \leq t_1^= \leq 11.333]) \vee [34.0 \leq t_8^+].$$

Note that this redescription was not returned previously since the beam search focused on better ones involving precipitation variables.

Filtering redundant redescriptions.

It is common to see a set of redescriptions that cover approximately the same area even if they have (somewhat) different sets of variables. Indeed, redescriptions belong to the family of local patterns, with each individual pattern independently describing a subset of the data. Mining local patterns typically returns redundant results that require filtering. In such cases, it is important to be able to recognize and remove redundant redescriptions, i.e. redescriptions that do not convey significant new information, lest the user be overwhelmed with the number of found redescriptions. Again, SIREN allows automatic filtering of redundant redescriptions. The user can select a redescription and ask SIREN either to filter out all redescriptions that are redundant with respect to the selected one, or to go through the whole list of redescriptions filtering out all redescriptions that are redundant with respect to some earlier-encountered (i.e. better) redescription. Naturally, the decisions made by SIREN can be reverted whenever the user wishes to.

For instance, the results returned during the extension mentioned previously may contain many redundant redescriptions found at different steps. We can easily sort them, e.g. by accuracy, select one of interest and filter all the following results redundant with respect to it.

Outputting the results.

Finally, SIREN facilitates the distribution of the results: redescriptions can be exported in easy-to-read format and the maps associated to redescriptions can be readily converted to publication-ready graphics.

4. REDESCRIPTION MINING

Redescription mining aims at simultaneously finding multiple descriptions of a subset of entities which is not previously specified. This is in contrast with other methods like Emerging Patterns Mining (EPM), Contrast Set Mining (CSM) and Subgroup Discovery (SD) (see [7] for a unifying survey) or general classification methods, where target subsets of entities are specified via labels. Currently, redescription mining is a purely descriptive approach, its predictive power remains to be explored. Since its introduction in [10] various algorithms have been proposed for Boolean redescription mining, based on approaches including decision trees [10, 5], co-clusters [8], and frequent itemsets [2].

At the core of SIREN is the REREMI redescription mining algorithm. This greedy algorithm uses an efficient on-the-fly discretization technique to extend redescription mining to categorical and numerical variables. Below, we give an outline of the concepts and algorithms involved. Full details can be found in the original publication [1].

We consider Boolean, categorical, and numerical variables, partitioned into two sets. Boolean variables can be interpreted as a truth value assignment in a natural way. For categorical and real-valued variables, truth value assignments are induced by relations $[v = c]$ and $[a \leq v \leq b]$, respectively, where c is some category and $[a, b]$ an interval. These truth assignments and their negations constitute *literals* which can be combined using the Boolean operators $\wedge$ (and) and $\vee$ (or) to form *queries*. Then, 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* 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.

We use a strategy similar to beam-search to explore the solution space. The basic idea is to construct queries bottom-up, starting from singleton redescriptions (i.e. both queries contain only one literal) and progressively extending them by appending operators and literals. After evaluating all possible one-step extensions, we select the best candidates and extend them in turn. This process stops when no new redescription can be generated.

5. IMPLEMENTATION DETAILS

SIREN and REREMI are implemented in Python. The interface is built with the `wxPython` Open Source GUI toolkit, ensuring cross-platform compatibility. The `matplotlib` library enables to generate high quality figures, seamlessly integrated in the interface. SIREN allows for simple editing of the redescriptions thanks to flexible parsing of different representations. It can handle any data provided in a compatible format.

The formulation of redescription mining presented here assumes that the describing variables are partitioned into two sets apriori, and looks for a pairs of queries over these two sets, respectively. However, this can be naturally adapted to settings with a single set of describing variables. One might then search for pairs of queries, with the constraint that the two subsets of variables appearing in the queries of any redescription be disjoint or enable the user to interactively determine the split between the variables.

6. CONCLUSIONS

We present SIREN, a tool for mining geospatial redescriptions. It enables users to interactively mine, edit and extend redescriptions. It also features visualization of the redescriptions on a map, a key toward interpreting the results of geospatial data mining.

We will give chance to the public to experiment with SIREN and together consider how this tool could be used to explore their own geospatial data and help answer their data analysis need.

7. REFERENCES

[1] E. Galbrun and P. Miettinen. From Black and White to Full Colour: Extending Redescription Mining Outside the Boolean World. In *SDM*, pages 546–557, 2011.

[2] A. Gallo, P. Miettinen, and H. Mannila. Finding subgroups having several descriptions: Algorithms for redescription mining. In *SDM*, pages 334–345, 2008.

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

[4] R. J. Hijmans, S. Cameron, L. Parra, P. Jones, and A. Jarvis. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.*, 25:1965–1978, 2005. `www.worldclim.org`.

[5] D. Kumar. *Redescription mining: Algorithms and applications in bioinformatics*. PhD thesis, Department of Computer Science, Virginia Tech, 2007.

[6] A. J. Mitchell-Jones, G. Amori, W. Bogdanowicz, B. Krystufek, P. Reijnders, F. Spitzenberger, M. Stubbe, J. Thissen, V. Vohralik, and J. Zima. *The atlas of European mammals*. Academic Press, London, 1999. `www.european-mammals.org`.

[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] L. Parida and N. Ramakrishnan. Redescription mining: Structure theory and algorithms. In *AAAI*, pages 837–844, 2005.

[9] 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.

[10] 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.

[11] J. Soberón and A. T. Peterson. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiv. Inform.*, 2(0), 2005.

[12] 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.