Siren Demo at ICDM 2015

Release 3.0.0

Esther Galbrun and Pauli Miettinen

November, 2015

Contents

Note: This webpage summarizes information about the demonstration of the *Siren* for mining redescriptions with trees, submitted to the 2015 IEEE International Conference on Data Mining, Atlantic City, NJ, USA on November 14-17, 2015.

Esther Galbrun and Pauli Miettinen. Mining predictive Redescriptions with Trees. Submitted to *ICDM*. 2015. Original paper.

More details can be found on the main *Siren* webpage or in the user guide.

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 their bioclimatic conditions and the fauna that inhabit them.

Siren is a tool for interactive mining and visualization of redescriptions. We integrated tree-based redescription mining algorithms, allowing to find redescriptions that generalize well.

1 Download

Siren is a multi-platform software implemented in Python.

- **Source code** (packaged using Python distutils): Siren (v3.0.0) .tar.gz

2 Data Formats

In *Siren*, data include:

- **Variables**: The variables describing the entities are divided in two sets. They can be of three types:
  1. Boolean,
  2. categorical,
  3. or real-valued.

Obviously, this is required.

- **Entities names**: Optional additional information, providing names for the entities.
- **Variable names**: Optional additional information, providing names for the variables.
- **Coordinates**: Optional location information, i.e. geographic coordinates of the entities. This makes the data geospatial.

Data can be imported to *Siren* via the interface menu *File* → *Import* → *Import Data*.

Data can be imported into *Siren* as CSV files. The program expects a pair of files, one for either side in character-separated values, as can be imported and exported to and from spreadsheet programms, for instance.

In particular, the data can stored as a table with one column for each variable and one row each entity. The first row can contain the names of the variables. The entities names can be included as columns named *ids*. Similarly the coordinates can be included as a pair of columns named *longitudes* and *latitudes*, respectively.

3 Mining redescriptions with trees

There are various strategies for mining redescriptions mining. We integrated tree-based algorithms to the *Siren* interface to allow mining redescriptions that generalize better, than, for instance redescriptions mine with the greedy *ReReMi* algorithm. For more details, check the *references* section.

3.1 CARTWheels (variant available in Siren)

The first algorithm introduced for redescription mining was actually based on alternating between constructing CARTs and hence was called the CARTWheels algorithm.

See the little slideshow below to understand how redescriptions are constructed with this approach and read the corresponding publication in the *references* section for more details.

3.2 Layered trees

An alternative method for constructing CARTs is to build them layer by layer, we call this method the *layered trees*.

3.3 Split trees

Finally the third method available in *Siren* construct queries by refining the CART branches separately, we call this method the *split trees*.

4 Sample Use-Cases

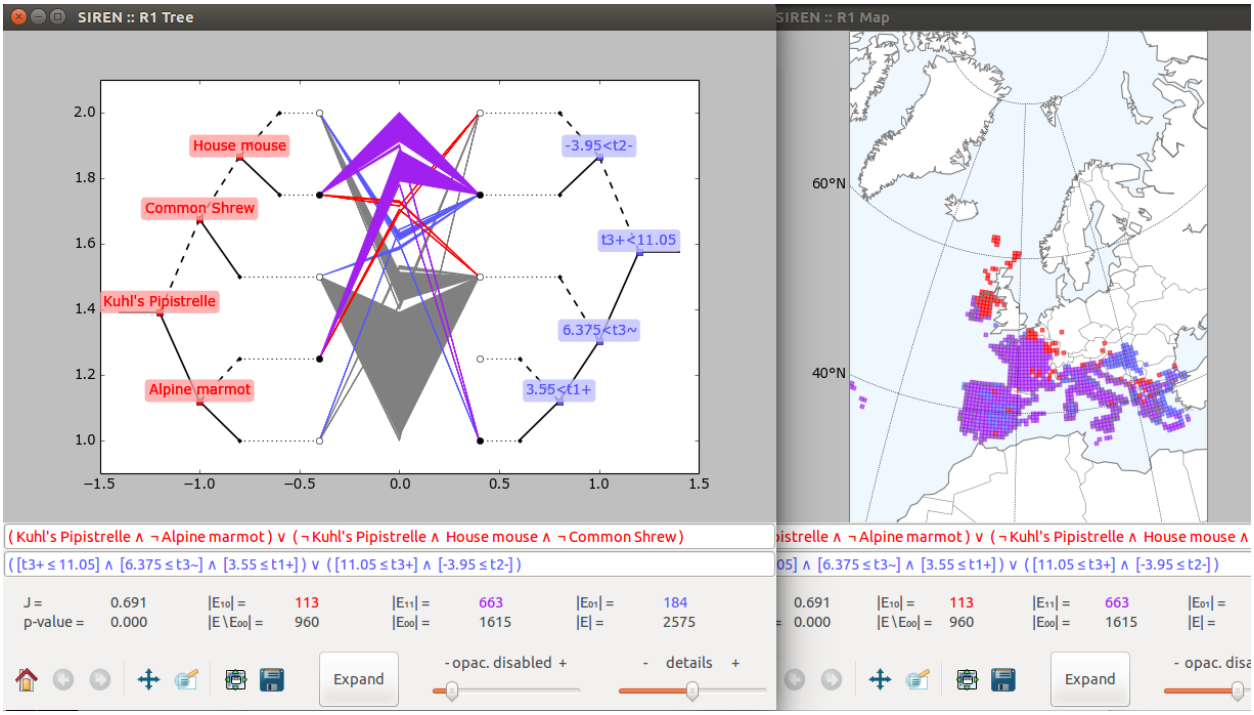
4.1 Finnish 2011 parliamentary elections

We provide a prepared dataset about the Finnish 2011 parliamentary elections. Get the data (non-geospatial), try out *Siren* and learn about the finnish political scene! (More details on the main webpage.)

To illustrate the use of *Siren*, we present example use-cases from different application domains.

4.2 Biological niche-finding

One use-case concerns niche-finding, i.e. the problem of finding species' bioclimatic envelope, an important task in biology.



5 Resources

References

[A] Esther Galbrun and Pauli Miettinen. Siren: an interactive tool for mining and visualizing geospatial redescriptions. In *KDD*, 1544–1547. ACM, 2012. Preprint, Poster.

[B] Esther Galbrun and Pauli Miettinen. From black and white to full color: extending redescription mining outside the Boolean world. *Statistical Analysis and Data Mining*, 5(4):284–303, 2012. Preprint.

[C] Esther Galbrun. Methods for Redescription Mining. PhD thesis, University of Helsinki, 2014. http://urn.fi/URN:ISBN:978-952-10-9431-6 .

[D] Tetiana Zinchenko. Redescription Mining Over non-Binary Data Sets Using Decision Trees. MSc thesis, University of Saarland and Max-Planck Institute for Informatics, 2014. http://www.mpi-inf.mpg.de/~pmiettin/papers/zinchenko15redescription.pdf .

[E] Naren Ramakrishnan, Deept Kumar, Bud Mishra, Malcolm Potts, and Richard F Helm. Turning CARTwheels: An Alternating Algorithm for Mining Redescriptions. In *KDD*, 266–275. ACM, 2004.