An Introduction to Redescription Mining

Tutorial

Esther Galbrun and Pauli Miettinen

April, 2017

Information about the tutorial on redescription mining at the SIAM International Conference on Data Mining (SDM 2017), taking place in Houston, TX, USA, on Thursday April 27, 2017.

The tutorial slides can be found here.

A biologist interested in bioclimatic habitats of species needs to find geographical areas that admit two characterizations, one in terms of their climatic profile and one in terms of the occupying species. For a political analyst, matching the personal profiles of an election's candidates to their viewpoints on various relevant issues might help cast light on the political scene and provide insight into the different candidates' positions.

These are just two examples of a general problem setting where we need to identify correspondences between data that have different nature (species vs. climate, personal profiles vs. political viewpoints). Beyond these two examples, redescription mining has diverse practical applications, from detecting criminal networks to optimizing circuit designs.

To identify the correspondences over binary data sets, Ramakrishnan et al. proposed redescription mining in 2004. Subsequent research has extended the problem formulation to more complex correspondences and data types, making it applicable to wide variety of data analysis tasks.

In this tutorial we will give an overview of redescription mining, from the intuition behind the concept and its links to existing data analysis techniques to more recent developments in algorithms and interactive mining techniques. We will also cover five areas where applications for redescription mining have been proposed. The tutorial will give the attendants knowledge of the state-of-the-art techniques in redescription mining and open problems in method development, as well as examples and information on how to apply redescription mining to real-world data analysis problems.

1 Overview

Redescription mining, introduced in 2004 by Ramakrishnan et al. *[Ram+04]*, aims at finding distinct common characterizations of the same objects. Like other multi-view analysis techniques, redescription mining is well suited to extract more coherent and relevant information, by exploiting different points of views on the same phenomena. This is an especially attractive feature as data grow increasingly diverse and heterogeneous. Recent advances in redescription mining algorithms and novel applications thereof — including the detection of criminal networks *[Wu+14]*, the optimization circuit designs *[Goe+10]* and the analysis political opinions *[GM16]* —have increased the interest in this relatively nascent subfield of data analysis.

We will start by giving a conceptual overview of redescription mining, from the intuition behind the concept and its links to existing data analysis techniques, before presenting the more recent technical developments in algorithms and visual interactive mining methods. Then, we will cover five areas where practical applications of redescription mining have been proposed. These case studies are meant to illustrate the wide variety of potential applications of redescription mining, as well as to highlight the strengths and weaknesses of the current methods in the different applications (see the *roadmap* for more details).

From this tutorial, the **data analysts** will gain an overview of the redescription mining framework, a necessary understanding of the state-of-the-art methods, and a good grasp of the type of problems redescription mining can be applied to, respectively. The **method developers**, on the other hand, will learn the context and problem definition, the stepping stones for building their own tools, and the practical requirements of redescription mining algorithms through concrete examples, respectively. More generally, the attendees of the tutorial will hear about redescription mining and related methods and will hence be able to use these techniques in their own research.

Redescription mining is an important data analysis technique that has been covered previously only once in a tutorial. While still arguably not wide-spread, redescription-based methods currently attract increasing amounts of research interest (see, e.g. *[GK14] [GM14] [ZGM15] [Wu+14]*). Furthermore, redescription mining techniques belong to the broader class of multi-view (or multi-modal) data analysis methods, which are becoming ever more important and relevant as the diversity and heterogeneity of available data grows.

Redescription mining employs machine learning techniques to automatically mine patterns from data, in order to support data exploration and knowledge discovery. Hence, it is particularly relevant to researchers from the data mining and analytics community who attend the SDM conference. This tutorial should help foster the research of redescription mining techniques and widen their use.

2 Audience

We expect the material to be interesting to a wide audience with diverse backgrounds, and to be accessible to anyone possessing basic knowledge of core data mining and machine learning techniques.

In this tutorial, you will learn what redescription mining is and what it is not. We will cover the two main views of redescription mining, association rule mining based and classification based, as well as the main related areas (subgroup discovery, subspace clustering, and multi-view data mining). You will learn about the state-of-the-art algorithms for finding redescriptions as well as interactive visualization techniques for exploring and interpreting them. The discussion will be illustrated with numerous examples throughout, and we will pay special attention to five practical applications of redescription mining in circuit design, bioinformatics, ecology, political sciences and data intelligence, respectively.

In short, you will learn about redescription mining and related methods and will hence be able to use these techniques in your own research.

This tutorial will last approximately 2 hours including. Questions will be welcomed throughout the tutorial and time will be left for discussion when wrapping up.

Welcome!

3 Tutors' bios

Pauli Miettinen is a senior researcher and head of the area Data Mining at the Databases and Information Systems department of the Max-Planck Institute for Informatics, Germany. He is also an Adjunct Professor (docent) of computer science at the University of Helsinki, Finland, where he previously worked in Prof. Heikki Mannila's group, and received his PhD in 2009. His main research interest is in Algorithmic Data Analysis. In particular, he has been working on matrix decompositions over non-standard algebras and their applications to data mining and on redescription mining. His research has resulted in numerous publications in top data mining venues, two best paper prices (PKDD '06 Best paper; PKDD '08 Best student paper), and an honorary mention at 2010 ACM SIGKDD Doctoral dissertation awards.

Pauli was the presenter of Decomposing Binary Matrices: Where Linear Algebra Meets Combinatorial Data Mining, a tutorial at ECML-PKDD 2012, and has given a number of invited talks and lectures.

http://people.mpi-inf.mpg.de/~pmiettin/

Esther Galbrun is a junior research scientist at INRIA Nancy–Grand Est, France. She was previously a postdoctoral researcher and part-time lecturer at the Computer Science department of Boston University, MA, USA after having obtained her PhD in 2014 from the Computer Science department at the University of Helsinki, Finland, on the topic of redescription mining.

https://members.loria.fr/EGalbrun/

4 Roadmap

1. **The Theory**
    - (a) **What is redescription mining?**
        - i. Introductory examples
        - ii. Definitions and problem formalization *[Gal13]*
        - iii. Sets of redescriptions *[DB11]*, *[KGM16]*
        - iv. The short history of redescriptions *[Ram+04] [PR05] [ZR05] [Kum+08]*, *[GMM08] [GM12a] [GK14] [ZGM15] [KGM16]*
    - (b) Related work *[Agg15] [Agr+98] [Coh+01] [BS04] [Gup+13] [KK14] [NLW09] [KZ09] [LFK08] [JMR08] [Ram+04] [Ume+09] [Wro97]*, *[Mie12] [ZGM15]*
2. The Techniques 1. Visualizing and interacting with redescriptions
    - (a) Visualizing objects with maps and projections *[GM12b]*
    - (b) Visualizing queries *[HS12] [Ins85]*, *[GM14]*
    - (c) Visualizing a set of redescriptions *[MS16]*
    - (a) **Algorithms for redescription mining**
        - i. Mining redescription with decision trees *[Bre+84] [Qui86] [Ram+04]*, *[ZGM15]*
        - ii. Combining frequent itemsets into redescriptions *[AS94] [GNDR13] [HPY00] [Zak+97] [ZH05] [ZR05] [ZZR06]*, *[GMM08]*
        - iii. Building redescriptions greedily *[GMM08] [GM12a]*
2. **The Practise**
    - (a) Bioinformatics: Biological pathways elucidation *[JMR08] [Kum07] [RZ09]*

(b) Ecology: Bioclimatic niche finding *[PD03] [PAS06] [SN09] [Thu+09]*, *[GM12a] [ZGM15]*

(c) Circuit Design: Sequential Equivalence Checking *[Goe+10]*

(d) Political sciences: Analyzing poll data *[GM16]*

(e) Data Intelligence: Storytelling *[Hos+12] [Kum+08] [Wu+14]*

- **Wrap-Up**
  - Summary

5 Resources

Slides

The tutorial slides can be found here.

Links

- Siren — Interactive and visual redescription mining, [webpage] [source code]

References

[Agg15] Charu C Aggarwal. *Data Mining*. Springer, 2015. Ch. 4.5.6. doi:10.1007/978-3-319-14142-8.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of 20th International Conference on Very Large Data Bases (VLDB'94)*, 487–499. 1994. URL: http://www.vldb.org/conf/1994/P487.PDF.

[Agr+98] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *ACM SIGMOD Record*, 27(2):94–105, 1998. doi:10.1145/276305.276314.

[BS04] Steffen Bickel and Tobias Scheffer. Multi-View Clustering. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04)*, 19–26. 2004. doi:10.1109/ICDM.2004.10095.

[Bre+84] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[Coh+01] Edith Cohen, Mayur Datar, Shinji Fujiwara, Aristides Gionis, Piotr Indyk, Rajeew Motwani, Jeffrey D Ullman, and Cheng Yang. Finding interesting associations without support pruning. *IEEE Transactions on Knowledge and Data Engineering*, 13(1):64–78, 2001. doi:10.1109/69.908981.

[DB11] Tijl De Bie. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *Data Mining and Knowledge Discovery*, 23(3):407–446, 2011. doi:10.1007/s10618-010-0209-3.

[Gal13] Esther Galbrun. *Methods for Redescription Mining*. PhD thesis, Department of Computer Science, University of Helsinki, 2013. URL: http://urn.fi/URN:ISBN:978-952-10-9431-6.

[GK14] Esther Galbrun and Angelika Kimmig. Finding relational redescriptions. *Machine Learning*, 96(3):225–248, 2014. doi:10.1007/s10994-013-5402-3.

[GM12a] 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. doi:10.1002/sam.11145.

[GM12b] Esther Galbrun and Pauli Miettinen. 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)*, 1544–1547. 2012. URL: http://adrem.ua.ac.be/iid2012/papers/galbrun_miettinen-visual_and_interactive_geospatial_redescription_mining.pdf.

[GM14] Esther Galbrun and Pauli Miettinen. Interactive Redescription Mining. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD'14)*, 1079–1082. 2014. doi:10.1145/2588555.2594520.

[GM16] Esther Galbrun and Pauli Miettinen. Analysing Political Opinions Using Redescription Mining. In *IEEE International Conference on Data Mining Workshops*, 422–427. 2016. URL: https://people.mpi-inf.mpg.de/~pmiettin/papers/galbrun16analysing.pdf.

[GMM08] Arianna Gallo, Pauli Miettinen, and Heikki Mannila. Finding Subgroups having Several Descriptions: Algorithms for Redescription Mining. In *Proceedings of the 8th SIAM International Conference on Data Mining (SDM'08)*, 334–345. 2008. doi:10.1137/1.9781611972788.30.

[Goe+10] Neha Goel, Michael S Hsiao, Naren Ramakrishnan, and Mohammed J Zaki. Mining Complex Boolean Expressions for Sequential Equivalence Checking. In *Proceedings of the 19th IEEE Asian Test Symposium (ATS'10)*, 442–447. 2010. doi:10.1109/ATS.2010.81.

[GNDR13] Tias Guns, Siegfried Nijssen, and Luc De Raedt. k-Pattern Set Mining under Constraints. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):402–418, 2013. doi:10.1109/TKDE.2011.204.

[Gup+13] Sunil Kumar Gupta, Dinh Phung, Brett Adams, and Svetha Venkatesh. Regularized nonnegative shared subspace learning. *Data Mining and Knowledge Discovery*, 26(1):57–97, 2013. doi:10.1007/s10618-011-0244-8.

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM Sigmod Record*, volume 29, 1–12. 2000. doi:10.1145/335191.335372.

[HS12] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *Communications of the ACM*, 55(4):45–54, 2012. doi:10.1145/2133806.2133821.

[Hos+12] M. Shahriar Hossain, Joseph Gresock, Yvette Edmonds, Richard Helm, Malcolm Potts, and Naren Ramakrishnan. Connecting the Dots between PubMed Abstracts. *PLoS ONE*, 7(1):1–23, 2012. doi:10.1371/journal.pone.0029509.

[Ins85] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985. doi:10.1007/BF01898350.

[JMR08] Ying Jin, T M Murali, and Naren Ramakrishnan. Compositional mining of multirelational biological datasets. *ACM Transactions on Knowledge Discovery from Data*, 2(1):2–35, 2008. doi:10.1145/1342320.1342322.

[KGM16] Janis Kalofolias, Esther Galbrun, and Pauli 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. URL: https://people.mpi-inf.mpg.de/~pmiettin/papers/kalofolias16from.pdf.

[KK14] Suleiman A Khan and Samuel 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)*, 656–671. 2014. doi:10.1007/978-3-662-44848-9_42.

[KZ09] Peer Kröger and Arthur Zimek. Subspace Clustering Techniques. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 2873–2875. 2009. doi:10.1007/978-0-387-39940-9_607.

[Kum07] Deept Kumar. *Redescription mining: Algorithms and applications in bioinformatics*. PhD thesis, Department of Computer Science, Virginia Polytechnic Institute and State University, 2007. URL: https://theses.lib.vt.edu/theses/available/etd-05032007-223232/unrestricted/deept_redescs.pdf.

[Kum+08] Deept Kumar, Naren Ramakrishnan, Richard F Helm, and Malcolm Potts. Algorithms for Storytelling. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):736–751, 2008. doi:10.1109/TKDE.2008.32.

[LFK08] Dennis Leman, Ad Feelders, and Arno J Knobbe. Exceptional Model Mining. In *Proceedings of the 2008 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'08)*, volume 5212, 1–16. 2008. doi:10.1007/978-3-540-87481-2_1.

[Mie12] Pauli Miettinen. On Finding Joint Subspace Boolean Matrix Factorizations. In *SIAM International Conference on Data Mining*, 954–965. 2012. doi:10.1137/1.9781611972825.82.

[MS16] Matej Mihelčić and Tomislav Šmuc. Interset: interactive redescription set exploration. In Toon Calders, Michelangelo Ceci, and Donato Malerba, editors, *Proceedings of the 19th International Conference on Discovery Science*, volume 9956 of Lecture Notes in Computer Science, 35–50. 2016. doi:10.1007/978-3-319-46307-0_3.

[NLW09] Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *Journal of the Machine Learning Research*, 10:377–403, 2009. doi:10.1145/1577069.1577083.

[PR05] Laxmi Parida and Naren 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)*, 837–844. 2005. URL: http://www.aaai.org/Library/AAAI/2005/aaai05-132.php.

[PD03] Richard G Pearson and Terence P Dawson. Predicting the Impacts of Climate Change on the Distribution of Species: Are Bioclimate Envelope Models Useful? *Global Ecology and Biogeography*, 12:361–371, 2003. doi:10.1046/j.1466-822X.2003.00042.x.

[PAS06] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3):231–259, 2006. doi:10.1016/j.ecolmodel.2005.03.026.

[Qui86] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):81–106, 1986. doi:10.1023/A:1022643204877.

[RZ09] Naren Ramakrishnan and Mohammed J Zaki. Redescription Mining and Applications in Bioinformatics. In J. Chen and S. Lonardi, editors, *Biological Data Mining*. 2009. URL: http://www.crcpress.com/product/isbn/9781420086843.

[Ram+04] Naren Ramakrishnan, Deept Kumar, Bud Mishra, Malcolm Potts, and Richard 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)*, 266–275. 2004. doi:10.1145/1014052.1014083.

[SN09] Jorge Soberón and Miguel Nakamura. Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences of the United States*, 106(Supplement 2):19644–19650, 2009. doi:10.1073/pnas.0901637106.

[Thu+09] Wilfried Thuiller, Bruno Lafourcade, Robin Engler, and Miguel B. Araújo. BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, 32(3):369–373, 2009. doi:10.1111/j.1600-0587.2008.05742.x.

[Ume+09] Lan Umek, Blaz Zupan, Marko Toplak, Annie Morin, Jean-Hugues Chauchat, Gregor Makovec, and Dragica 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)*, volume 5651, 265–274. 2009. doi:10.1007/978-3-642-02976-9_39.

[Wro97] Stefan 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)*, volume 1263, 78–87. 1997. doi:10.1007/3-540-63223-9_108.

[Wu+14] Hao Wu, Jilles Vreeken, Nikolaj Tatti, and Naren Ramakrishnan. Uncovering the plot: detecting surprising coalitions of entities in multi-relational schemas. *Data Mining and Knowledge Discovery*, 28(5-6):1398–1428, 2014. doi:10.1007/s10618-014-0370-1.

[ZH05] Mohammed J. Zaki and Ching-Jui Hsiao. Efficient Algorithms for Mining Closed Itemsets and their Lattice Structure. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):462–478, 2005. doi:10.1109/69.846291.

[ZR05] Mohammed J Zaki and Naren Ramakrishnan. Reasoning About Sets Using Redescription Mining. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'05)*, 364–373. 2005. doi:10.1145/1081870.1081912.

[Zak+97] Mohammed J Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li. New Algorithms for Fast Discovery of Association Rules. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'97)*, 283–286. 1997. URL: http://www.aaai.org/Library/KDD/1997/kdd97-060.php.

[ZZR06] Lizhuang Zhao, Mohammed J Zaki, and Naren Ramakrishnan. BLOSOM: A framework for mining arbitrary boolean expressions. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, 827–832. 2006. doi:10.1145/1150402.1150511.

[ZGM15] Tetiana Zinchenko, Esther Galbrun, and Pauli Miettinen. Mining predictive redescriptions with trees. In *IEEE International Conference on Data Mining Workshops*, 1672–1675. 2015. doi:10.1109/ICDMW.2015.123.