HAL
open science

The Coron System

Mehdi Kaytoue, Florent Marcuola, Amedeo Napoli, Laszlo Szathmary, Jean Villerd

To cite this version:

Mehdi Kaytoue, Florent Marcuola, Amedeo Napoli, Laszlo Szathmary, Jean Villerd. The Coron System. 8th International Conference on Formal Concept Analsis - ICFCA 2010, Mar 2010, Agadir, Morocco. pp.55–58. inria-00600232

HAL Id: inria-00600232

https://inria.hal.science/inria-00600232

Submitted on 23 Nov 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Coron System

Mehdi Kaytoue[1], Florent Marcuola[1], Amedeo Napoli[1], Laszlo Szathmary[2], and Jean Villerd[1]

[1] Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)
Campus Scientifique – BP 239 – 54506 Vandœuvre-lès-Nancy Cedex (France)
{kaytouem, marcuolf, napoli, villerd}@loria.fr

[2] Département d'Informatique – Université du Québec à Montréal (UQAM)
C.P. 8888 – Succ. Centre-Ville, Montréal H3C 3P8 (Canada)
Szathmary.L@gmail.com

Abstract. Coron is a domain and platform independent, multi-purposed data mining toolkit, which incorporates not only a rich collection of data mining algorithms, but also allows a number of auxiliary operations. To the best of our knowledge, a data mining toolkit designed specifically for itemset extraction and association rule generation like Coron does not exist elsewhere. Coron also provides support for preparing and filtering data, and for interpreting the extracted units of knowledge.

Key words: knowledge discovery, data mining, itemset extraction, association rules generation, rare item problem

1 System Overview

Born for a particular need in a cohort study [1], Coron is now a framework of knowledge discovery in databases on its own, used in several application domains, e.g. [4–6]. Intended to an educational and scientific usage, the Coron system is articulated into several modules for preparing and mining binary data, and filtering and interpreting the extracted units. Thus, from binary data (possibly obtained from a discretization procedure), Coron allows one to extract itemsets (frequent, closed, generators, etc.) and then to generate association rules (non-redundant, informative, etc.). Building concept lattices is also possible. The system includes many classical algorithms of the literature, but also others that are specific to Coron [9–11]. The software is freely available at `http://coron.loria.fr`. Mainly written in Java, Coron is compatible with the Unix, Mac and Windows operating systems and is of command-line usage.

2 A Global Data Mining Methodology

The methodology was initially designed for mining biological cohorts, but it is generalizable to any kind of database. It is important to notice that the whole process is guided by an expert, who is a specialist of the domain related to the database. His role may be crucial, especially for selecting the data and for

interpreting the extracted units, in order to fully turn them into knowledge units. In our case, the extracted knowledge units are mainly association rules. At the present time, finding association rules is one of the most important tasks in data mining. Association rules allow one to reveal "hidden" relationships in a dataset. Finding association rules requires first the extraction of frequent itemsets.

The methodology consists of the following steps: Definition of the study framework; Iterative step: data preparation and cleaning, pre-processing step, processing step, post-processing step; Validation of the results and Generation of new research hypotheses; Feedback on the experiment. The life-cycle of the methodology is shown in Figure 1. Coron is designed to satisfy the present methodology and offers all the tools that are necessary for its application in a single platform.

Pre-processing. These modules propose several tools for manipulating and formatting large data. The data are described by binary tables in a simple text-file format: some individuals in lines possess or not some properties in column. The main possible operations are: (i) discretization of numerical data, (ii) conversion of different file formats, (iii) creation of the complement of the binary table, and (iv) other projection operations such as transposition of the table.

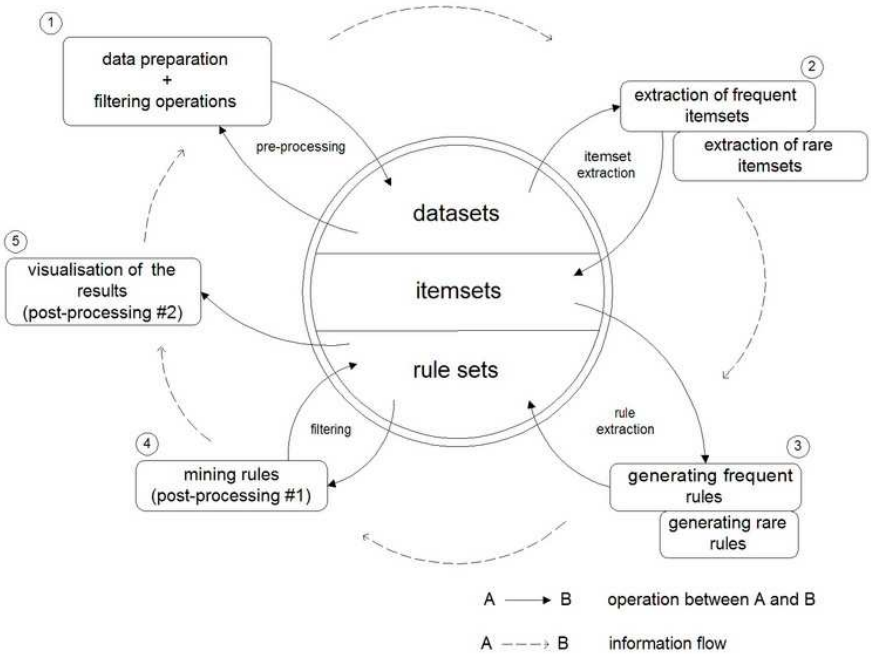

Fig. 1. Architecture of the Coron System

Data mining. Extracting itemsets and association rules is a very popular task in data mining. Concept lattices are mathematical structures supported by a rich and well established formalism, namely, Formal Concept Analysis [13]. A concept lattice is represented by a diagram giving nice visualization of classes of objects of a domain. Thus, the data mining modules of the Coron System offer the following possibilities:

- Itemset extraction: frequent, closed, rare, generators, etc. This task is performed by a large collection of algorithms based on different search strategies (depth-first, level-wise, etc.).
- Association rules generation: frequent, rare, closed, informative, minimal non-redundant, Duquenne-Guigues basis, etc. These rules are given with a set of measures such as support, confidence, lift, conviction, etc.
- Concept lattice construction.

Post-processing. Extracted units from the data mining step may be very numerous, and hide some units of higher interest. Thus, Coron proposes some filtering operations that should be done in interaction with a domain expert. The analyst may filter rules w.r.t. the length of its components, and/or the presence of a given property. He may also retain the k best extracted units w.r.t. a measure of interest. It is also possible to color some properties of a list of association rules.

Toolbox. Finally, auxiliary modules allow one to visualize equivalence classes of itemsets, randomly generate binary data, etc.

3 Applications

Coron has been used for the following tasks: extraction of knowledge of adaptation in case-based reasoning [4], gene expression data analysis [5, 12], information retrieval [7], recommendations for internet advertisement [6], biological data integration [8], and finally, cohort studies [1].

4 Work in Progress

Currently, we are studying how to integrate Coron in platforms using graphical data-flows, such as Knime [2], whose popularity is increasing (`http://www.knime.org`). This would allow Coron to interact with many other useful tools, most importantly avoiding a command-line usage. Also, other tools will be integrated in Coron to consider complex data, mainly numerical, see e.g. [12]. Finally, we have recently set up a forum to gather questions, comments and suggestions from Coron users (`http://coron.loria.fr/forum/`).

In this paper, we have given a brief overview of the Coron System. For more details, please refer to the project's website at `http://coron.loria.fr`.

Acknowledgements

The authors would like to thank the following persons for their participation in the development of Coron: F. Collignon, B. Ducatel, S. Maumus, P. Petronin, T. Bouton, A. Knobloch, N. Sonntag, Y. Toussaint.

References

1. L. Szathmary, S. Maumus, P. Petronin, Y. Toussaint and A. Napoli, Vers l'extraction de motifs rares. Actes de *Extraction et Gestion de connaissances (EGC), RNTI-E-6, Cépaduès-Éditions Toulouse*, pages 499–510, 2006
2. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Koetter, T. Meinl, P. Ohl, C. Sieb, and B. Wiswedel, Knime: The Konstanz Information Miner. Demonstration at *Knowledge Discovery in Databases (KDD)*, 2006
3. L. Szathmary, A. Napoli and P. Valtchev, Towards Rare Itemset Mining, *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 305–312, 2007
4. M. d'Aquin, F. Badra, S. Lafrogne, J. Lieber, A. Napoli and L. Szathmary, Case Base Mining for Adaptation Knowledge Acquisition. Proc. of the *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 750–755, 2007
5. M. Kaytoue, S. Duplessis and A. Napoli, Using Formal Concept Analysis for the Extraction of Groups of Co-expressed Genes. Proc. of the *International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences (MCO), CCIS, Springer*, 439–449, 2008
6. D. I. Ignatov and S. O. Kuznetsov, Concept-based Recommendations for Internet Advertisement. Proc. of the *Concept Lattices and Their Applications (CLA)*, pages 157–166, 2008
7. E. Nauer and Y. Toussaint, Classification dynamique par treillis de concepts pour la recherche d'information sur le web. Actes de *5ème conférence de recherche en information et applications (CORIA)*, pages 71–86, 2008
8. A. Coulet, M. Smaïl-Tabbone, P. Benlian, A. Napoli and M.-D. Devignes, Ontology-guided data preparation for discovering genotype-phenotype relationships. *BMC Bioinformatics*, Vol. 9, 2008
9. L. Szathmary, A. Napoli, S.O. Kuznetsov, ZART: A Multifunctional Itemset Mining Algorithm, Proc. of the *5th Intl. Conf. on Concept Lattices and Their Applications (CLA)*, pages 26–37, 2007
10. L. Szathmary, P. Valtchev, A. Napoli and R. Godin, Constructing Iceberg Lattices from Frequent Closures Using Generators, Proc. of the *International Conference on Discovery Science (DS), LNCS 5255, Springer*, pages 136–147, 2008
11. L. Szathmary, P. Valtchev, A. Napoli and R. Godin, Efficient Vertical Mining of Frequent Closures and Generators, Proc. of the *International Symposium on Intelligent Data Analysis (IDA), LNCS, Springer*, pages 393–404, 2009
12. M. Kaytoue, S. Duplessis, S. O. Kuznetsov and A. Napoli, Two FCA-Based Methods for Mining Gene Expression Data, Proc. of the *International Conference on Formal Concept Analysis (ICFCA), LNCS 5548, Springer*, pages 251–266, 2009
13. B. Ganter and R. Wille, Formal Concept Analysis, Mathematical Foundations, Springer, 1999