New Applications of Formal Concept Analysis: A Need for Original Pattern Domains

Jean-François Boulicaut

INSA Lyon, LIRIS CNRS UMR5205
F-69621 Villeurbanne cedex, France
Jean-Francois.Boulicaut@insa-lyon.fr

Abstract. We survey the results obtained by our research group (joint work with Jérémy Besson and Loïc Cerf, Kim-Ngan T. Nguyen, Marc Plantevit, and Céline Robardet) concerning the design of pattern domains to support knowledge discovery and information retrieval in arbitrary n-ary relations. Our contribution is related to Formal Concept Analysis and its recent developments in direction of, for instance, Triadic Concept Analysis. We focus on a real data mining perspective. It means that we need for both the design of scalable constraint-based mining algorithms and fault-tolerant approaches to support the discovery of relevant patterns from noisy data.

1 Extended abstract

The Formal Concept Analysis framework (FCA) has been studied for about three decades [13]. Given a binary relation, we may consider FCA as the computation and then the exploitation of a collection of closed patterns, the so-called formal concepts that are organized within a lattice structure. FCA supports knowledge discovery processes from such relations and many application domains have been considered. It includes applications to more or less simple information retrieval tasks (see, e.g., [5, 21]).

Nowadays, we have to face with more and more large but also structured types of data like, for instance, (collections) of graphs or information networks. New challenges have appeared such that pattern discovery methods have to be revisited. One important direction of research concerns the extension of FCA-based techniques for different types of data (e.g., numerical matrices, collections of strings). For instance, this is currently studied thanks to the concept of pattern structure [17, 16]. Also, Triadic Concept Analysis that concerns Boolean cube data analysis has been formalized in [18] and several algorithms have been proposed to discover patterns in such ternary relations (see, e.g., the computation of closed patterns [14, 15] or implications [12]). For instance, it can be applied to the discovery of conceptual structures in folksonomies that are ternary relations $Users \times Resources \times Tags$.

During the last decade, our research group[1] has been working on various evolutions of FCA where (a) datasets are arbitrary n-ary relations, (b) computed

[1] liris.cnrs.fr/equipes?id=46

patterns are not only closed but must also satisfy other user-defined primitive constraints, and (c) some fault-tolerance is provided.

Following the guidelines of inductive querying and constraint-based data mining [4, 11], we have been designing new pattern domains. The methodology is as follows.

Given a data type, we have to define pattern languages and measures that denote properties of patterns within the data. Then, we carefully design the primitive constraints that will be combined to support the declarative specification of both objective and subjective interestingness. Once declarative specifications are available - the so-called inductive queries - we must provide algorithms that compute the solution patterns. A major issue is to identify the constraint properties and the enumeration strategies that enable to compute correct and complete answers in practical cases. For this, generic algorithms can be designed: no specific combination of primitive constraint is expected but safe pruning theorems can be based on the constraint properties. Notice that it is generally possible to design more efficient ad-hoc algorithms when considering fixed forms of constraints.

In our 2008 survey [3], we were considering a constraint-based perspective on actionable formal concept mining from large binary relations. As a result, we were discussing the use of primitive constraints to compute more relevant formal concepts, for instance large-enough ones [2] but also some generalizations that provide fault-tolerance [1]. A few years later, it is now possible to discuss such issues in the enlarged setting of arbitrary n-ary relations. Therefore, we can consider (a) our generic algorithm that mines set patterns and exploits the large class of piecewise (anti-)monotonic constraints [7, 8], (b) its extension towards fault-tolerant pattern discovery by means of a correct and complete strategy [6] or an heuristic one [9]. We also studied a multidimensional association rule mining framework [19] that is based on closed pattern post-processing. Among others, promising though preliminary applications to dynamic relational graph analysis have been investigated [10, 20].

References

1. J. Besson, R. Pensa, C. Robardet, and J.-F. Boulicaut. Constraint-based mining of fault-tolerant patterns from boolean data. In KDID'05 Revised Selected and Invited Papers, volume 3933 of LNCS, pages 55–71. Springer, 2005.
2. J. Besson, C. Robardet, J.-F. Boulicaut, and S. Rome. Constraint-based formal concept mining and its application to microarray data analysis. Intell. Data Anal., 9(1):59–82, 2005.
3. J.-F. Boulicaut and J. Besson. Actionability and formal concepts: A data mining perspective. In Proc. ICFCA, volume 4933 of LNCS, pages 14–31. Springer, 2008.
4. J.-F. Boulicaut, L. D. Raedt, and H. Mannila, editors. Constraint-Based Mining and Inductive Databases, volume 3848 of LNCS. Springer, 2005.
5. C. Carpineto and G. Romano. Using concept lattices for text retrieval and mining. In Proc. ICFCA, volume 3626 of LNCS, pages 161–179. Springer, 2005.
6. L. Cerf, J. Besson, K.-N. Nguyen, and J.-F. Boulicaut. Closed and noise-tolerant patterns in n-ary relations. Data Min. Knowl. Discov., 26(3):574–619, 2013.

7. L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut. Data peeler: Contraint-based closed pattern mining in n-ary relations. In Proc. SIAM DM, pages 37–48, 2008.
8. L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut. Closed patterns meet n-ary relations. ACM Transactions on KDD, 3(1), 2009.
9. L. Cerf, P.-N. Mougel, and J.-F. Boulicaut. Agglomerating local patterns hierarchically with ALPHA. In Proc. ACM CIKM, pages 1753–1756, 2009.
10. L. Cerf, T. B. N. Nguyen, and J.-F. Boulicaut. Mining constrained cross-graph cliques in dynamic networks. In Inductive Databases and Queries: Constraint-Based Data Mining, pages 201–230. Springer, 2010.
11. S. Dzeroski, B. Goethals, and P. Panov, editors. Inductive Databases and Queries: Constraint-Based Data Mining. Springer, 2010.
12. B. Ganter and S. A. Obiedkov. Implications in triadic formal contexts. In Proc. ICCS, volume 3127 of LNCS, pages 186–195. Springer, 2004.
13. B. Ganter, G. Stumme, and R. Wille, editors. Formal Concept Analysis, Foundations and Applications, volume 3626 of LNCS. springer, 2005.
14. R. Jaschke, A. Hotho, C. Schmitz, B. Ganter, and G. Stumme. Trias–an algorithm for mining iceberg tri-lattices. In Proc. IEEE ICDM, pages 907–911, 2006.
15. L. Ji, K.-L. Tan, and A. K. H. Tung. Mining frequent closed cubes in 3D data sets. In Proc. VLDB, pages 811–822, 2006.
16. M. Kaytoue, S. O. Kuznetsov, and A. Napoli. Revisiting numerical pattern mining with formal concept analysis. In Proc. IJCAI, pages 1342–1347, 2011.
17. S. O. Kuznetsov. Pattern structures for analyzing complex data. In Proc. RSFDGrC, volume 5908 of LNCS, pages 33–44. Springer, 2009.
18. F. Lehmann and R. Wille. A triadic approach to formal concept analysis. In Proc. ICCS, volume 954 of LNCS, pages 32–43. Springer, 1995.
19. K.-N. Nguyen, L. Cerf, M. Plantevit, and J.-F. Boulicaut. Multidimensional association rules in boolean tensors. In Proc. SIAM DM, pages 570–581, 2011.
20. K.-N. Nguyen, L. Cerf, M. Plantevit, and J.-F. Boulicaut. Discovering descriptive rules in relational dynamic graphs. Intell. Data Anal., 17(1):49–69, 2013.
21. J. Poelmans, D. I. Ignatov, S. Viaene, G. Dedene, and S. O. Kuznetsov. Text mining scientific papers: A survey on fca-based information retrieval research. In Proc. ICDM, volume 7377 of LNCS, pages 273–287. Springer, 2012.