

# An AI Planning System for Data Cleaning

Roberto Boselli<sup>1,2</sup>, Mirko Cesarini<sup>1,2</sup>, Fabio Mercurio<sup>1,2</sup> (✉),  
and Mario Mezzanzanica<sup>1,2</sup>

<sup>1</sup> Department of Statistics and Quantitative Methods,  
University of Milano-Bicocca, Milan, Italy  
`fabio.mercurio@unimib.it`

<sup>2</sup> CRISP Research Centre, University of Milano-Bicocca, Milan, Italy

**Abstract.** Data Cleaning represents a crucial and error prone activity in KDD that might have unpredictable effects on data analytics, affecting the believability of the whole KDD process. In this paper we describe how a bridge between AI Planning and Data Quality communities has been made, by expressing both the data quality and cleaning tasks in terms of AI planning. We also report a real-life application of our approach.

**Keywords:** AI planning · Data quality · Data cleaning · ETL

## 1 Introduction and Motivation

A challenging issue in data quality is to automatically check the quality of a source dataset and then to identify cleaning activities, namely a sequence of actions able to cleanse a dirty dataset. Data quality is a domain-dependent concept, usually defined as “fitness for use”, thus reaching a satisfying level of data quality strongly depends on the analysis purposes. Focusing on *consistency*, which can be seen as “the violation of semantic rules defined over a set of data items” [1], the state-of-the-art solutions mainly rely on functional dependencies (FDs) and their variants, that are powerful in specifying integrity constraints. Consistency requirements are usually defined on either a single tuple, two tuples, or a set of tuples [4]. Though the first two kind of constraints can be modelled through FDs, the latter one requires reasoning with a (finite but variable in length) set of data items (e.g., time-related data), and this makes the use of FD-based approaches uneffective (see, e.g., [4, 10]). This is the case of *logitudinal data* (aka *historical or time-series data*), which provide knowledge about a given subject, object or phenomena observed at multiple sampled time points. In addition, it is well known that FDs are expressive enough to model static constraints, which evaluate the current state of the database, but they do not take into account how the database state has evolved over time [3]. Furthermore, though FDs enable the detection of errors, they cannot be used as guidance to fix them [9].

In such a context graphs or tree formalisms are deemed also appropriate to model the *expected* data behaviour, that formalises how the data should evolve

over time for being considered as consistent, and this makes the exploration-based technique (as AI Planning) a good candidate for the data quality task. The idea that underlies our work is to cast the problem of checking the consistency of a set of data items as a planning problem. This, in turn, allows using off-the-shelf AI planning tools to perform two separated tasks: (i) to catch inconsistencies and (ii) to synthesise a sequence of actions able to cleanse any (modelled) inconsistencies found in the data. In this paper we summarise results from our recent works on data consistency checking [15] and cleaning [2, 14].

*AI Planning at a Glance.* Planning in Artificial Intelligence is about the decision making performed by computer programs when trying to achieve some goal. It requires to synthesise a sequence of actions that will transform a system configuration, step by step, into the desired one (i.e., the goal state). Roughly, planning requires two main elements: (i) *the domain*, i.e., a set of states of the environment  $S$  together with the set of actions  $A$  specifying the transitions between these states; (ii) *the problem*, that consists of the set of facts whose composition determine an initial state  $s_0 \in S$  of the environment, and a set of facts  $G \subseteq S$  that models the goals of the planning task. A solution (aka *plan*) is a bounded sequence of actions  $a_1, \dots, a_n$  that can be applied to reach a goal configuration. Planning formalisms are expressive enough to model complex temporal constraints, then a cleaning approach based on AI planning might allow domain experts to concentrate on *what* quality constraints have to be modelled rather than on *how* to check them. Recently, AI Planning contributed to the trace alignment problem in the context of Business Process Modelling [5].

## 2 A Data Cleaning Approach Framed Within KDD

Our approach requires to map a sequence of events as actions of a planning domain, so that AI planning algorithms can be exploited to find inconsistencies and to fix them. Intuitively, let us consider an events sequence  $\epsilon = e_0, e_2, \dots, e_{n-1}$ . Each event  $e_i$  will contain a number of observation variables whose evaluation determines a snapshot of the subject's *state*<sup>1</sup> at time point  $i$ , namely  $s_i$ . Then, the evaluation of any further event  $e_{i+1}$  might change the value of one or more state variables of  $s_i$ , generating a new state  $s_{i+1}$ .

We encode the expected subjects' behaviour (the so-called *consistency model*) as a transition system. A consistent trajectory represents a sequence of events that does not violate any consistency constraints. Given a  $\epsilon$  event sequence as input, the planner deterministically determines a trajectory  $\pi = s_0 e_0 s_1 \dots s_{n-1} e_{n-1} s_n$  on the finite state system explored (i.e., a *plan*) where each state  $s_{i+1}$  results by applying event  $e_i$  on  $s_i$ . Once a model describing the evolution of an event sequence has been defined, we detect quality issues by solving a planning problem where a consistency violation is the goal condition. If a plan is found by a planning system, the event sequence is marked as inconsistent in the original data quality problem. Our system works in three steps (Fig. 1).

<sup>1</sup> A value assignment to a set of finite-domain state variables.



*Real-life Application*<sup>3</sup>. Our approach has been applied to the *mandatory communication*<sup>4</sup> domain, that models labour market data of Italian citizens at regional level. Here, inconsistencies represent career transitions not permitted by the Italian Labour Law. Thanks to our approach, we synthesised both the Universal Checker and Cleanser for the domain (i.e., 342 distinct inconsistencies found and up to 3 cleaning action sequence synthesised for each). The system has been employed within the KDD process that analysed the real career sequences of 214,432 citizens composed of 1,248,751 mandatory notifications. For details about the quality assessment see [15] whilst for cleaning details see [14].

### 3 Concluding Remarks

We presented a general approach that expresses Data quality and cleaning tasks in terms of AI Planning problem, connecting two distinct research areas. Our approach has been formalised and fully-implemented on top of the UPMurphi planner, and applied to a real-life example analysing and cleaning million records concerning labour market movements of Italian citizens.

We are working on (i) including machine-learning algorithms to identify the *most suited* cleaning action, and (ii) applying our approach to build training sets for data cleaning tools based on machine-learning (e.g., [11]).

### References

1. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications. Springer, New York (2006)
2. Boselli, R., Cesarini, M., Mercorio, F., Mezzanica, M.: Planning meets data cleansing. In: The 24th ICAPS. AAAI Press (2014)
3. Chomicki, J.: Efficient checking of temporal integrity constraints using bounded history encoding. ACM Trans. Database Syst. (TODS) **20**(2), 149–186 (1995)
4. Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A.K., Ilyas, I.F., Ouzzani, M., Tang, N.: NADEEF: a commodity data cleaning system. In: SIGMOD (2013)
5. De Giacomo, G., Maggi, F.M., Marrella, A., Patrizi, F.: On the disruptive effectiveness of automated planning for LTLf-based trace alignment. In: AAAI (2017)
6. Della Penna, G., Intrigila, B., Magazzeni, D., Mercorio, F.: UPMurphi: a tool for universal planning on PDDL+ problems. In: The 19th ICAPS, pp. 106–113 (2009)
7. Della Penna, G., Intrigila, B., Magazzeni, D., Mercorio, F.: A PDDL+ benchmark problem: the batch chemical plant. In: ICAPS, pp. 222–224. AAAI Press (2010)
8. Della Penna, G., Magazzeni, D., Mercorio, F.: A universal planning system for hybrid domains. Appl. Intell. **36**(4), 932–959 (2012)
9. Fan, W., Li, J., Ma, S., Tang, N., Yu, W.: Towards certain fixes with editing rules and master data. Proc. VLDB Endowment **3**(1–2), 173–184 (2010)

<sup>3</sup> This work was partially supported within a Research Project granted by the CRISP Research Centre and Arifl Agency (Regional Agency for Education and Labour.

<sup>4</sup> The Italian Ministry of Labour and Welfare: Annual report about the CO system, available at <http://goo.gl/XdALYd> last accessed may 2017.

10. Hao, S., Tang, N., Li, G., He, J., Ta, N., Feng, J.: A novel cost-based model for data repairing. *IEEE Trans. Knowl. Data Eng.* **29**(4) (2017)
11. Krishnan, S., Wang, J., Wu, E., Franklin, M.J., Goldberg, K.: ActiveClean: Interactive data cleaning while learning convex loss models. arXiv preprint [arXiv:1601.03797](https://arxiv.org/abs/1601.03797) (2016)
12. Mercorio, F.: Model checking for universal planning in deterministic and non-deterministic domains. *AI Commun.* **26**(2), 257–259 (2013)
13. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: Data quality sensitivity analysis on aggregate indicators. In: *DATA*, pp. 97–108 (2012)
14. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: A model-based approach for developing data cleansing solutions. *ACM J. Data Inf. Q.* **5**(4), 1–28 (2015)
15. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: A model-based evaluation of data quality activities in KDD. *Inf. Process. Manag.* **51**(2), 144–166 (2015)