26/09/2022

Talk 2022

T1. Let me first thank the organizers for this workshop … This idea to have a 20-years-after milestone is interesting. I had the feeling that something special was planned at KDID 2022 … Well it was the case and, so far, I enjoyed all the talks. As some of you know, this is probably my last talk in public. I prepared it with that in mind and I'll use some time for memories and acknowledgements.

T2. KDID has been a dissemination tool for European projects about inductive databases. Late nineties, more and more ad-hoc data mining algorithms were proposed and we were lacking of generic approaches. A database perspective was sketched in early papers by A. Siebes, T. Imielinski and H. Mannila. In 1997, I asked to join Heikki Mannila's group in Helsinki for a sabbatical year. My goal was to learn data mining, and, hopefully, to contribute at some time. 25 years after, I can say "I love it when a plan comes together!"

T3. So, late October 1997, I arrived in Helsinki. Let me provide a pretty fair transcription of the first meeting I had with Heikki Mannila.
> What are your plans?
I would like to investigate the concept of inductive database to support real database reverse engineering ... May be using Inductive Logic Programming ...
> Hum ... Real databases ... May be too difficult to do something useful
Whoops ...
> Reverse engineering ... Difficult to do something useful
Ouch ...
> Inductive Logic Programming ... Why ...
Arghh ...
> Let us look at some open problems in Boolean data
It has been the starting point of a quite nice adventure … Notice from that transcription that I adopted, almost immediately, the speaking conciseness from Finland ☺

T4. The concept of inductive databases can be illustrated in my "collector" drawing (a few variants exist but its graphical charter is protected ☺). The user can query the data, the patterns and the models that hold in the data. KDD processes then become querying processes. A major issue about patterns and models is that an intent can be queried such that, at some time, it will be needed to compute them. In the simple formalization of inductive queries by H. Toivonen and H. Mannila, we see the components of a pattern domain, i.e., a data type, a pattern language and a collection of primitive constraints that can be combined.

T5-T6. The first basic research project on inductive databases cInQ has involved 6 partners between 2001 and 2004 (Lyon, Torino, Milano, Freiburg, Helsinki and Ljubljana). At this time, all the researchers from Helsinki were affiliated with its Nokia Research Center. The second FET project IQ, its follow-up, has involved 6 partners (Ljubljana, Lyon, Leuven, Antwerp, Helsinki and Aberystwyth … always hard to pronounce) between 2005 and 2008. At this time, the group lead by Luc de Raedt had left Freiburg for Leuven.

T7-T8. So KDID has been a dissemination tool for cInQ and IQ. We organized 5 workshops co-located with ECML PKDD from 2002 until 2006. 3 of them have given rise to post-workshop books. The project partners also edited 3 books that contain tens of papers on inductive databases and constraint-based data mining. Let us consider a subjective selection of results from cInQ and IQ.

T9. We have been working on many local pattern domains and algorithms for computing inductive queries on them. Doing so, we have contributed to constraint-based mining of association rules, bi-sets or formal concepts, episode rules, substrings or sequential patterns. Less known pattern domains have been investigated like 1st order sequences in Freiburg and polynomial equations in Ljubljana. In Lyon we have studied n-ary relation mining. It has been applied, among others, to pattern discovery in relational dynamic graphs.

T10. Let us consider a Boolean cube that encodes the directional edges between a fixed set of vertices at successive time stamps. A maximal clique is in fact a straightforward extension of formal concepts in Boolean matrices: we look here for combinatorial maximal cubes of true values. We can then exploit efficiently other constraints like, e.g., the clique constraint, a large-enough volume constraint or a contiguity constraint for the temporal dimension.

T11. Assume that data encode strong interaction between individuals during ECML PKDD from 2002 (Helsinki) to 2008 (Antwerp). We can discover such an almost preserved maximal clique. Can we interpret this? OK, these are smart data miners from Europe … but we know other ones … What do they share? Five of them were invited in 2002 at the ESF workshop on Pattern Detection organized by David Hand in London. Also, they have been together at the Dagstuhl seminar on Local Pattern Detection in 2004. The remaining member, Céline, was at this time the rising star in Lyon. This group has also addressed the compulsive beer open question. We all heard about suspected associations between milk, diapers and beer but … a rigorous study remains missing. We started such a task by developing an expertise about beers in every country where ECML PKDD was organized ☺

T12. An obvious outcome has been the study of condensed representations for 0/1 data but not only. The initial motivation was to get condensed representations of frequent sets and $\gamma/2$ adequate representations of the data for a frequency threshold γ. Closed sets became immediately a good candidate but others have been proposed. Some of these studied condensed representations have been also relevant per se and not only as a container for frequencies. For instance, interesting subsets of frequent and valid association rules can be based on δ-free sets and their closures.

T13. For many local patterns, we can work on generic exhaustive strategies based on primitive constraint properties (typically monotonicity properties). Interesting transformations of inductive queries have been studied in Lyon, Freiburg and Helsinki. Another axis has been the cross-fertilization with constraint programming. It turns out that we can model data mining tasks in terms of constraints for which efficient solvers have been optimized for decades. After results for itemset mining, it has been extended to other tasks and this is currently a well identified research axis.

T14. It was one of the main objectives of the IQ project to focus on constraint-based mining for global patterns or models. Several methods have been proposed for constraint-based mining of clusters and co-clusters but also decision trees. We studied various ways to build classifiers based on collections of local patterns, e.g., associative classifiers or the use of δ-free itemsets as computed features to support difficult classification tasks.

T15. When we started the cInQ project, we had the University of Torino and the Politecnico di Milano in the consortium. They described their MINE RULE operator earlier in 1996. This was a query language to support association rule mining only. An interesting outcome of IQ that implements the inductive database concept stricto sensu has been the MINING VIEW system prototyped in Antwerp and Leuven. They show how to model data mining output by relational tables that can then be queried thanks to SQL views. Scenarios that involve decision trees and association rules have been prototyped.

T16. Most of the research that has been done within the cInQ and IQ projects have been targeted towards methods and not applications. In other terms, what is called an application for us has been often the, possibly sophisticated effort to provide empirical feedback on our new algorithms. We however report here examples of application-oriented contributions by colleagues in Lyon, Aberystwyth and Leuven, Torino and Milano and finally the Nokia Research Center in Helsinki.

T17. My last research project DUF 4.0 (2018-2021) has concerned recipe optimization in urban vertical farms. In such engineering systems, the artificial climate management unit provides many ways to optimize the production of plants where they are needed, e.g., in the center of cities. A recipe is a sequence of desired values for many measures like, e.g., the number of seeds per pot, the led frequencies, the temperature, the hygrometry, the strength of wind, the amount of nutriments, etc. For each recipe, we can have one or several target attributes that can be computed at crop time, e.g., a yield, a gustative score or an energy cost.

T18. We developed an algorithm that computes optimal subgroups in numerical data with one numerical target. We show that it is possible to compute such a subgroup and its description from a set of used recipes. Then, using the optimal subgroup description, we can propose new recipes, apply them and enter a virtuous circle. It has been validated partially on real data (urban farm prototype from a start-up in Lyon) and with a sophisticated scenario based on a crop simulation environment.

T19. Multi-target optimization appears quite useful here. For instance, we may consider target attributes like yield and cost. It is interesting to consider the Pareto front of the recipes w.r.t. these attributes. We developed an algorithm that discovers a description of a subgroup of recipes whose Pareto front approximates well the data one. Here again, we can use it to design new recipes, apply them and enter a virtuous circle. We have a pretty long and well written research report that provide many details about this and a sophisticated scenario based on the crop simulation environment.

T20. Time to conclude. Do we have recent publications addressing explicitly inductive databases … not really. Where are the Inductive Database Management Systems? … I do not know … nowhere. The short-term amazing pressure on the applications of Data Science (often prototyped thanks to workflows and/or Python scripts) does not help to promote research on such concepts. It happens ;-) It does not mean that we will never see convincing algebras or inductive database management systems … However, it already happened that promising and intrinsically smart concepts and tools disappear from research and/or industrial agendas (e.g., logic programming or deductive databases). It makes me think that it may be good that I am retiring … I have been working on compiler compilers … the domain has disappeared … I have been working on Logic Programming … The domain has disappeared … I have been working on Inductive Databases … The domain sounds a bit moribund … ☺

T21. The research on constraint-based data mining is closely related to inductive databases and it remains a major sub-area of data mining. Many researchers are working on « complex » data mining with new pattern domains. We have nice challenges for heterogeneous data mining, e.g., multi-graphs whose vertices and/or edge attributes can contain measures, texts, images or videos. I like the Exceptional Model Mining research domain and its incredible potential for useful knowledge discovery. It should become a major sub-area. Finally, I am not aware of compiling approaches for sequences of data mining tasks like in workflows. It could be motivated by both runtime efficiency and the potential for static analysis.

T22. Are we interested in Data Mining and Machine Learning or Data Mining for Machine Learning? For instance, when my department decided a 2 years ago to open a new course on Machine Learning and was looking for modules that should disappear … I had to argue that it would be an error to replace my own data mining course – assumed backward-looking - by a machine learning one … convincing that both are needed for nowadays engineers. Data Mining is often used during the earlier steps of a KDD process where the objective is to understand the data but it can also be used to discover knowledge by means of descriptive methods. It is needed to develop further descriptive approaches and it includes pattern discovery both before Machine Learning and after Machine Learning. Indeed, it has become a topic for Explainable Artificial Intelligence. In that direction, let me point out the recent work by colleagues at LIRIS about pattern discovery on the layers of trained Graph Neural Networks.

T23. The future for me … after years of hard work … more time in Marseillan for the family … i.e., my wife Françoise, Claire my daughter, Lucas my grand-son and his sister (a question of days).

T24. Let me comment a couple of photos about the places or "things" that have been important for me during my career. Important places … Lyon in France, Helsinki in Finland, Pisa in Italy, Antwerp in Belgium, the amazing room for Ph. D defenses in Utrecht (Netherlands), Ljubljana in Slovenia, Praha in the Czeck Republik, Barcelona in Spain, Berlin, Hinterzarten but also the Dagstuhl castle in Germany, London in the United Kingdom, Montreal in Canada and Japan. Due to you and the memories from KDID 2022, I also add Grenoble … You can see also a few "objects" … like bottles of beers, wine and saké, but also the PACE flag from ECML PKDD 2004 in the Pisa venue or my collection of Arto Paasilinna books … in French ☺.

Let me finally thanks some colleagues … at a work life scale … (from T25 to T30).