Theoretical Challenges in Knowledge Discovery in Big Data
a logic reasoning and a graph theoretical point of view

Pavel Surynek[*], Petra Surynková[**]
Charles University Prague, Faculty of Mathematics and Physics
[*] *Malostranské náměstí 25, 118 00 Praha, Czech Republic*
[**] *Sokolovská 83,186 75 Praha 8, Czech Republic*
{pavel.surynek, petra.surynkova}@mff.cuni.cz

Keywords: big data, data analysis, logic reasoning, graph theory, graph drawing, propositional satisfiability

Abstract: This paper addresses a problem of knowledge discovery in big data from the point of view of theoretical computer science. Contemporary characterization of big data is often preoccupied by its volume, velocity of change, and variety that causes technical difficulties to handle the data efficiently while theoretical challenges that are offered by big data are neglected at the same time. Contrary to this preoccupation with technical issues, we would like to discuss more theoretical issues focused on the goal briefly expressed as what be understood from big data by imitating human like reasoning through logic and algorithmic means. The ultimate goal marked out in this paper is to develop an automation of the reasoning process that can manipulate and understand data in volumes that is beyond human abilities and to investigate if substantially different patterns appear in big data than in small data.

1 INTRODUCTION

Contrary to contemporary understanding of big data (Laney, 2012), which focuses on technological managing of difficulties arising from its still increasing volume, velocity of change, and growing variety of sources, we would like to discuss issues connected with a question what can be learned from big data by computational techniques. That is, we would like to discuss the big data challenge more from the point of view of theoretical computer science and artificial intelligence (Russell & Norvig, 2009). To simplify the situation we need to look aside from technical issues for now. Regarding mentioned technical difficulties known as 'V's (*velocity*, *volume*, *variety*, *value*, *veracity*) let us settle with any solution that allows us to access data in a convenient way and do not address this issue any further.

What we consider more exciting about big data than managing their volumes and what is currently addressed insufficiently is automated interpretation of data and automated learning from them. This issue has not yet been addressed in any significant extent and even the terminology for describing problems we would like to discuss is lacking. Many techniques already exist, but they are scattered in many other areas and not focused on big data directly. It is one of the goals of this paper is to point out techniques that can be employed in big data processing. The next goal is to show problems that arise in big data and that can be studied theoretically. Terms of knowledge discovery and reasoning in big data are closest titles for problems we consider interesting from theoretical point of view that we would like to discuss. However, these titles should be understood as working ones.

We will pick several concrete theoretical problems in big data to describe current big data challenges concretely. A solving approach, that should be considered and that is promising for a thorough investigation, is suggested for each of the mentioned problems. We will show big data problems from the perspective theoretical fields of *mathematical logic* and *graph theory*. All the concrete problems are put into context of related works and background; thus this work may serve a brief survey as well.

1.1 Big Data (vs. Small Data)

The core inspiration for the discussion is a question how to imitate human like reasoning over small data, which are met by humans every day, by algorithmic

techniques. Such automation allows applying of the imitated reasoning on large amounts of data that is beyond human capabilities. The adopted prerequisite in this study is that humans use logical reasoning.

Consider for example human ability of driving a car. The driver perceives visual, audio, and tactile data, which he quickly processes to make efficient decisions such as to accelerate. The situation with the human driver can be regarded as a small data world (even though one may object that the amount of processed data is still big). A corresponding big data world may take into consideration all the cars in a city or even a country at once. The outcome of the automated reasoning over such big data situation may be a prediction or a decision that for example prevents a traffic jam.

Analogical examples can be found in how humans extract knowledge from textual data, how they combine facts to answer questions, or how they understand social relations to join profitable coalitions. Successful automation in such cases brings possibility of building knowledge from whole libraries of text or predicting large-scale social trends based on understanding relations of large communities.

A very interesting question is that if substantially different patterns appear in big data from those that appear in small data. That is, if quantity of data leads to a quality that cannot be observed in small data. We consider this question as ultimate goal of effort in understanding big data through logical, algorithmic, and graph theoretical means.

2 CHALLENGES IN BIG DATA

The basic challenge in big data can be characterized as *knowledge discovery*. Extracting knowledge from big data is a prerequisite for making automated reasoning over the data.

One of the concrete approaches to knowledge discovery in data from the point of view of theoretical computer science is to try to find a (logical) theory that represents data in a compact form. The intuition behind this approach is that the compact form of the representation inherently induces certain kind of understanding, explanation, or structural insight – without understanding and discovering intrinsic rules in the data set, the compactness would be impossible (see Figure 1 for illustration of this intuition).

If data are interpreted as facts or statements, the aim is to find a theory in which these facts or statements are valid (Dwe Battista et al., 1998). Formally said, the set of models of the theory would be equal to the represented data set (Hodges, 1993). Regarding equality between both sets, one does not need to be that strict. Certain level of approximation of the data set by the theory should be also considered. Availability of such a theory then allows further decision making like checking of validity of new propositions, checking of consistency of a set of statements, finding the smallest set of statements that lead to a contradiction with the theory, and many other decisions known from logic reasoning.

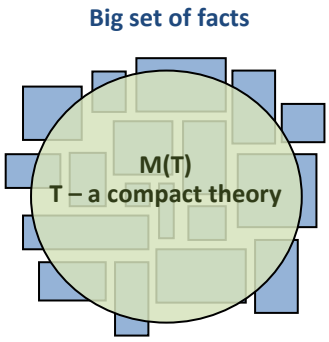

Figure 1. *Data representation as a set of models of a logical theory T.* The set of models of the theory M(T) approximates the input big data. The theory should be small through which certain level of understanding or explanation of data can be obtained.

Considering data in their big amounts may lead to finding novel understandings and interpretations. Historically, logic theories were used as formalizations of human reasoning hence it is quite natural to apply automated logic reasoning to process large amounts of data, which is consistent with suggested original inspiration. A question how to find logical representations of data sets algorithmically is discussed in following sections. Several approaches that should be further elaborated are suggested.

2.1 Compact data representation for their better understanding

Assume that the input data has the form of a set of logical statements. An important pool of techniques that should be considered consists of compression techniques for such a set of logical statements. Compression is regarded as a tool for discovering compact explanation of the given set of data.

The first step is to model (logical) data as a set of vectors over the propositional or multiple-value domain. Then it is almost immediate idea to investigate possibilities of their representation using some

existing concept such as *binary decision diagrams* (BDDs) (Akers, 1978) or *multi-value decision diagrams* (MDDs) (Miller & Drechsler, 1998). Although mentioned concepts are primarily intended as compact representations of the set of models of a certain formula (Rice, 2008) the huge source of results in this topic can be utilized in big data research as well.

Techniques for constructing decision diagrams themselves may be enriched within big data research as we expect big data to offer different challenges. The major difference can be observed in the fact that the whole process in data representation is reversed if it is compared with representation of models of a formula. Normally, the set of models of the formula is found and captured explicitly by the decision diagram. In data representation on the other hand, we start with explicit data set and through the intermediate step consisting of a decision diagram we want to understand the data. That is, to find a formula or a set of formulae (a theory), in which data are valid.

Decision diagrams are not the only concepts for data representation and compression. Another interesting method for data compression is represented by *matrix factorization* (Koren et al., 2009) and *matrix sketching* (Liberty, 2013). The former one has been recently successfully employed in *recommender systems* (Ricci et al., 2011). These methods compress large sparse matrices by representing them as products of smaller matrices where certain tolerance is given to the accuracy of the represented matrix. They are particularly attractive for their ability to discover hidden interpretations of data, which has been demonstrated by discovering hidden features in case of recommender systems.

Another interesting way to discover knowledge is to extract information from some kind of computational model or classifier known from *machine learning* (Mitchell, 1997) such as *neural network* (Zhang, 2000) or *Bayesian network* (Pearl, 1988). This approach has been already successfully used in many variations. The most notable example of knowledge extraction from the computational model has been done with *neural network* from which logic programs were extracted (Lehmann et al., 2010).

It seems to be promising to continue in research in knowledge extraction from computational models in the context of big data. A suitable computational model can be learned from the input training data and then further processed. The advantage here is that many efficient training algorithms for constructing computational models from training data already exist – in case of neural networks, *back-propagation* algorithm (Rumelhart et al., 1986) exists to name some. However, the large size of training data must be considered at this stage when dealing with big data. As existing training algorithms are not primarily designed for big data, the situation may lead to developing novel training methods in order to manage learning stage in acceptable time. In any case, it is expected that the outcome of the process will be a computational model that represents training data in the compact form. Then information can be extracted from the computational model.

The concrete way how to extract information is subject of further research and cannot be answered within this discussion. Nevertheless, it is assumed that the target of information extraction will be certain logic theory. No less important advantage of machine learning techniques is that they are typically robust with respect to inconsistencies and inaccuracies in the training data sets. Inconsistency represents an important issue in big data collected from some real-life source (Huang et al., 2013). Thus, the burden of dealing with data inconsistencies can be partly passed on learning process of the given computational model.

2.2 Deciding big data problems in description logic through SAT solving

A well-developed framework that provides rich description concepts and variety of decision methods is represented by *description logic* (DL) (Knorr et al., 2011). Currently, description logic is often used as a knowledge representation tool in *semantic web* and *bioinformatics* as it excels in expressing statements about individuals from some domain (such as medicine). Decision problems in description logic include testing if certain individual belong to given category or whether given individuals are bound together by a relation. Generally, decision problems in description logic can be regarded as more advanced and more complex variant of database querying (Bienvenu et al., 2013). Again, it is very interesting to use DL for representation of big data sets; and to apply DL reasoning and decision procedures to derive meaningful facts from the data set.

DL itself is extremely broad topic, thus a realistic attitude towards DL in perspective of big data is rather to just pick decision procedures suitable for application in big data reasoning and eventually to adapt and improve these procedures. The problematic point of application of DL with respect to big data is complexity of its decision procedures (Lutz, 2002). Although problems in DL are mostly decidable, the complexity of associated decision proce-

dures is often too high to be considered scalable – usually decision problems are *PSPACE-complete* (Baader et al., 2008), which practically means intractability especially when the input is big data. There are certain restrictions such as Horn-DL (Krötzsch et al., 2013) in which some decision problems are easier, that is in P, which makes it an interesting option for reasoning with big data.

However, the drawback of easier decision procedures may be that information discovered by such a procedure is invaluable. Usually worthwhile knowledge or information is difficult to discover, therefore a decision procedure that employs search to certain extent is needed.

A possible way to tackle difficulty of deciding in DL is to investigate possibilities of applying modern SAT solvers (Eén & Sörensson, 2004), (van Maaren & Franco, 2013), which are famous for their efficiency in searching for a solution, which is in their case a valuation of propositional variables that satisfies the given propositional formula. To make application of SAT solvers possible on knowledge discovery in big data an encoding of associated decision problems as propositional satisfiability is needed. Some advances in modeling decision problems in DL as SAT has been already made (Sebastiani & Vescovi, 2009). This recent progress is focused on modeling classical queries of DL in propositional satisfiability. A promising research direction is to find how to enrich this approach with the aspect of large amounts data translated to propositional statements or facts. It is known that state-of-the-art SAT solvers can find satisfying valuation of formulae containing up to millions of variables – such formulae often appear in hardware verification. Big data may become another domain where SAT solvers are successfully applied as such data are expected to contain regular patterns similarly as it is in the case of hardware verification formulae.

As it has been mentioned, data collection may contain inaccuracies and inconsistencies, which may compromise the application of crisp reasoning methods like SAT, which does distinguish only two cases – satisfiable and unsatisfiable but nothing in between. The situation is different in *MaxSAT* (Argelich et al., 2008), (Battiti & Protasi, 1998) where it is tried to satisfy maximum number of clauses in the given propositional formula. Such kind of optimization is worth considering for modeling problems in knowledge discovery. For example finding maximally consistent subset of statements in big data set is a viable candidate for such modeling.

2.3 Visualization and analysis of big data supported by graph theoretical techniques

Lot of understanding of not only big data but also data generally can be bolstered by visualization. The fascinating point with data visualization is that it combines computer graphics and combinatorial problem solving which represents a nice opportunity for cross-fertilization.

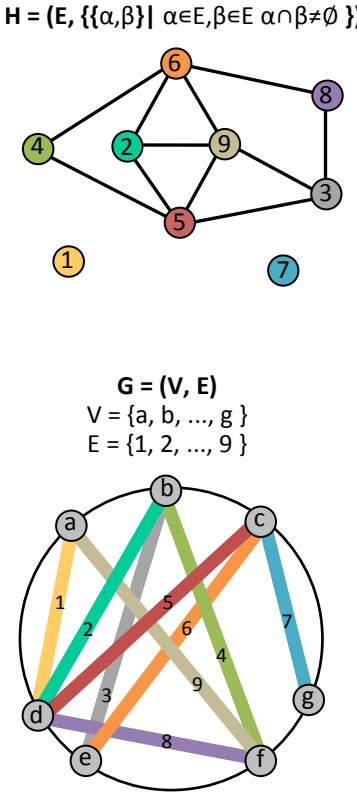

Figure 2. *An interpretation of linked data as a chordal graph.* The left graph H is a representation of a small case of linked data. The right graph is an alternative representation of links by intersections between chords of the cycle. One of the aims of the project is to find suitable visualizations through various types of intersection graphs for big cases of linked data.

Here, WE would like to discuss more the combinatorial aspect of data visualization. Data has the form of relations in many cases (Hitzler & Janowicz, 2013) where the relation says if a given tuple of objects are related or not. Special kind of relation is a binary relation, which considers ordered pairs of objects. This is the most frequent relation and the most studied one. Binary relation can be also under-

stood as a link between given objects. Therefore, data consisting of such relations are called *linked data* and their processing is called *linked data analysis* (Joshi et al., 2013).

The data set containing binary relations can be abstracted as a directed graph where objects are represented as vertices and binary relations between objects are represented as directed edges (or links; usually depicted like arrows). Having a graph or a big graph in the case of processing big data, we immediately face the problem how to visualize it or draw it. A classical problem of drawing a graph in plane where edges do not intersect, which gave rise to the definition of *planar graphs* for which it is possible (Di Battista et al., 1998), can serve as a starting point. One can also optimize the number of edge intersections to obtain best possible drawing of a graph.

Huge amounts of results exist in graph visualization. There is even a conference dealing solely with combinatorial aspects of graph visualization (Di Battista et al., 2014). The challenge connected with big data is that considered graphs are extremely large. Thus even polynomial time algorithms in other areas considered as efficient may be prohibitively slow in the case of big data. Hence, methods that process tasks connected with visualization in linear time should be in focus.

Many efficient (linear-time) visualization techniques for graphs can be found in so-called *intersection graphs* (Golumbic, 1980). Edges in intersection graphs are defined as intersection between some objects such as intervals or chords within a cycle (see Figure 2 for illustration of a chordal graph). The important feature of intersection graphs is that certain visualization is captured directly by the definition. Special objects that do intersect give the resulting graph special properties. Typically combinatorial problems, which are difficult in general graphs such as determining the *chromatic number* or the *clique number*, are easy in some cases of intersection graphs. It is worth studying if intersection graphs can be derived from big data and if this knowledge can be utilized in efficient data visualization.

Generally, we consider visualization as a tool to discover new hypotheses about visualized concepts. Data visualization has been applied with non-trivial success to find new ways how to optimize solution of problems in theoretical robotics (Surynek, 2011). Therefore, we expect lot from visualization in big data analysis.

3 EXPECTED PROGRESS

We would like summarize progress that we expect in mentioned aspects of big data processing in this section.

It is expected to find concepts that allow understanding of (big) datasets through compression. A variant of decision diagram that allows compact representation of dataset from which important features of data can be extracted (similarly as it is done in case of decision trees) is an expectable result for instance. Fundamental properties of suggested concepts are expected to be described and evaluated by means of theoretical computer science.

Efficient encodings of decision problems from big data into propositional satisfiability are expected to be found for example. This is connected with identifying interesting decidable problems in big data. An extension of existent applications of SAT in description logic to big data issues is expected. Again, fundamental properties should be described and theoretically as well as experimentally evaluated. Overcoming the crisp reasoning in *SAT* paradigm to make it suitable for supposedly inaccurate data possibly by shifting to *MaxSAT* would be valuable.

There are two expectable types of contributions regarding data visualization. The first should be development of a collection of supportive software prototypes to enable observation of big data through innovative visualizations. The supporting role of such software consists in helping to understand what is important in big data, which can show promising research directions. The second type of outcome is represented by fundamental combinatorial findings that allow visualization. Discovery of suitable graph drawing techniques is expected.

In my opinion, the ultimate type of contribution to the research in big data would be a discovery of a pattern that structurally distinguishes big data collection from the small one. That is, a pattern that is not observable in the small scale.

4 CONCLUSIONS

My goal has been to indentify several challenges in big data research from the point of view of theoretical computer science and artificial intelligence.

The paramount problem in big data we have focused on is knowledge discovery. Several particular problems related to knowledge discovery are identified and approaches how to address them are

discussed. We identify three challenges and their prospective solutions:

(i) Knowledge discovery through compression of the set of facts is suggested to be solved by using decision diagrams like BDD or MDD.

(ii) Decision problems in big data are suggested to be solved by translating them to description logic. Possible solution to tackle the complexity of associated decision procedures is application modern SAT solvers.

(iii) Finally, we see a great potential in solving combinatorial problems related to big data visualization if regarded as graphs.

The paper also represents a brief survey of theoretically oriented works applicable in knowledge discovery.

ACKNOWLEDGEMENTS

This research is work is supported by the Czech Science Foundation under the contract number GAP103/10/1287.

REFERENCES

Akers, S.: *Binary decision diagrams.* IEEE Transactions on Computers, Vol. 27(6), 509–516, IEEE Press, 1978.

Argelich, J., Li, C.-M., Manya F., Planes, J.: *The First and Second Max-SAT Evaluations.* Journal on Satisfiability, Vol. 4, 251-278, IOS Press, 2008.

Baader, F., Hladik, J., Peñaloza, R.: *Automata can show PSpace results for Description Logics.* Information and Computation, Special Issue: LATA 2007, 206(9-10):1045-1056, Elsevier, 2008.

Battiti, R., Protasi, M.: *Handbook of Combinatorial Optimization*. Kluwer, 1998.

Bienvenu, M., Ortiz, M., Simkus, M.: *Conjunctive Regular Path Queries in Lightweight Description Logics.* Proceedings of IJCAI 2013, IJCAI/AAAI, 2013.

Di Battista, G., Eades, P., Tamassia, R., Tollis, I. G.: *Graph Drawing: Algorithms for the Visualization of Graphs.* Prentice-Hall, 1998.

Di Battista, G., Tamassia, R., Tollis, I. G.: *International Symposium on Graph Drawing.* Web Page, http://www.graphdrawing.org/, 2014, [accessed in February 2014].

Eén, N., Sörensson, N.: *An Extensible SAT-solver*. Proceedings of SAT 2003, LNCS 2919, 502-518, Springer Verlag, 2004.

Golumbic, M. C.: *Algorithmic Graph Theory and Perfect Graphs.* Academic Press, 1980.

Hitzler, P., Janowicz, K.: *Linked Data, Big Data, and the 4th Paradigm.* Semantic Web, 4(3), 233-235, IOS Press, 2013.

Hitzler, P., Krötzsch, M., Rudolph, S.: *Foundations of Semantic Web Technologies.* Textbooks in Computing, Chapman and Hall/CRC Press, 2009.

Hodges, W.: *Model Theory.* Cambridge University Press, 1993.

Huang, S., Li, Q., Hitzler, P.: *Reasoning with Inconsistencies in Hybrid MKNF Knowledge Bases.* Logic Journal of the IGPL 21 (2), 263-290, Oxford University Press, 2013.

Joshi, A., Hitzler, P., Dong, G.: *Logical Linked Data Compression.* Proceedings of ESWC 2013, LNCS 7882, 170-184, Springer Verlag, 2013.

Knorr, M., Alferes, J. J., Hitzler, P.: *Local Closed-World Reasoning with Description Logics under the Well-founded Semantics.* Artificial Intelligence 175 (9-10), 1528-1554, Elsevier, 2011.

Koren, Y., Bell, R. M., Volinsky, C.: *Matrix Factorization Techniques for Recommender Systems.* IEEE Computer, Vol. 42 (8), 30-37, IEEE Press, 2009.

Krötzsch, M., Rudolph, S., Hitzler, P.: *Complexities of Horn Description Logics.* ACM Transactions on Computational Logic, Vol. 14 (1), ACM Press, 2013.

Laney, D.: *The Importance of 'Big Data': A Definition*. Gartner, 2012.

Lehmann, J., Bader, S., Hitzler, P.: *Extracting Reduced Logic Programs from Artificial Neural Networks.* Applied Intelligence, Vol. 32(3), 249-266, Springer Verlag, 2010.

Liberty, E.: *Simple and deterministic matrix sketching*. Proceedings KDD 2013, 581-588, ACM Press, 2013.

Lutz, C. *The complexity of Description Logics with concrete domains.* PhD Thesis, LuFG Theoretical Computer Science, RWTH Aachen, Germany, 2002.

Maaren, H. van, Franco, J.: *The international SAT Competitions.* Competition web page, http://www.satcompetition.org/, 2013, [accessed in February 2014].

Miller, D. M., Drechsler, R.: *Implementing a multiple-valued decision diagram package.* Proceedings of ISMVL 1998, 52-57, IEEE Press, 1998.

Mitchell, T.: *Machine Learning.* McGraw Hill, 1997.

Pearl, J.: *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.

Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (editors): *Recommender Systems Handbook.* Springer Verlag, 2011.

Rice, M.: *A Survey of Static Variable Ordering Heuristics for Efficient.* BDD/MDD Construction. University of California, 2008.

Rumelhart, D. E., Hinton, G. E., Williams, R. J.: *Learning representations by back-propagating errors.* Nature, Vol. 323 (6088): 533–536, Nature Publishing, 1986.

Russell, S. and Norvig, P.: *Artificial Intelligence: A Modern Approach.* Prentice Hall, 2009.

Sebastiani, R., Vescovi, M.: *Automated Reasoning in Modal and Description Logics via SAT Encoding: the Case Study of K(m)/ALC-Satisfiability.* J. AI Res., Vol. 35: 343-389, AAAI Press, 2009.

Surynek, P.: *Redundancy Elimination in Highly Parallel Solutions of Motion Coordination Problems.* Proceedings of ICTAI 2011, 701-708, IEEE Press, 2011.

Zhang, G. P.: *Neural Networks for Classification: A Survey.* IEEE Transactions on Systems, Man, and Cybernetics—part C: Applications and Reviews, Vol. 30 (4), IEEE Press, 2000.