Inducing models of vehicular traffic complex vague concepts by interaction with domain experts

Paweł Gora
Faculty of Mathematics, Computer Science and Mechanics
University of Warsaw
Warsaw, Poland
Email: pawelg@mimuw.edu.pl

Piotr Wasilewski
Faculty of Mathematics, Computer Science and Mechanics
University of Warsaw
Warsaw, Poland
Email: piotr@mimuw.edu.pl

***Abstract*—In the paper, we outline our research on obtaining domain knowledge related to vehicular traffic in cities using interaction with experts. The goal of acquiring such knowledge is to construct hierarchical domain oriented classifiers for approximation of complex vague concepts related to the road traffic. Interaction with experts in construction of hierarchical classifiers is supported by the software for agent-based simulation of vehicular traffic in cities, Traffic Simulation Framework, developed by the first author.**

***Keywords*-vehicular traffic; interaction with expert; complex vague concepts; perception based computing;**

I. INTRODUCTION

Vehicular traffic in cities is a complex phenomenon, which has a significant impact on environment and life of many people. Understanding the phenomenon and learning how to control it is a very important task.

One of the main objectives of our research is to detect traffic jam patterns from low level data using domain knowledge. We propose to support the searching process by interaction with domain experts ([2]). This can be done by acquiring from experts the relevant concepts, e.g. *traffic jam*, *traffic congestion*, *traffic jam formation*, and next by making it "understandable" to the system using classifiers. The key issue here is how to dialogize with experts.

In this research, we focus on a single basic traffic concept - *traffic congestion on a single crossroad* - and we elaborate methods for approximating this concept from sensory data. Sensory data come from simulating traffic using the Traffic Simulation Framework software [6], [7], [8]. Data from the software may slightly differ from real-world traffic data (which are very difficult to obtain), but are confirmed to be quite realistic [9], enough to conduct our research and get meaningful results. These data will be used to construct hierarchical classifiers based on rough set methods [2], [16], [22], which will approximate the concept of a traffic jam on a single crossroad. The concept is complex, vague and semantically distant from sensory data, so it is difficult to construct such classifiers explicitly. Classifiers constructed using universal methods (independent on domain knowledge) were not able to approximate such complex traffic concepts with satisfactory accuracy. It is necessary to construct domain oriented classifiers [2]. However, it is not clear how to obtain domain knowledge related to such complex phenomenon as traffic jam and it motivates research presented in the paper.

The paper has the following organization. In Section II we present the idea of interaction with domain experts and explain why it is important in the contemporary machine learning and data mining, particularly in acquiring domain knowledge about complex processes as vehicular traffic in cities. In Section III, we argue that vehicular traffic should be considered as a complex system and its understanding and modeling is a difficult task. Section IV outlines past approaches to traffic modeling, recent approaches based on probabilistic cellular automatons and the model developed and implemented by the first author of the paper. Section V presents our methodology in details: the procedure of dialogizing with domain experts, design of our experiments and expert decisions evaluation. In this section, we also present values of parameters that are used in our traffic simulations.

II. INTERACTION WITH DOMAIN EXPERTS

Contemporary machine learning faces a couple of big challenges. One of them is the problem of data mining (DM) and knowledge discovery in databases (KDD) with dynamically evolving complex data (e.g. stream data sources, sensory data). Another challenge for machine learning is a growth of size and complexity of data sources (e.g. Web sources, neuro-imaging data, data from network interactions). These challenges, in particular, discovery of complex concepts, hardly can be met by classical methods [19]. They can be met by KDD systems dialogizing with experts or users (e.g. interview with V. Vapnik [29]) or by adaptive learning systems changing themselves during the learning process as the response to evolving data. Another challenge comes from a field of multi-agent systems. Behavior steering and coordination of multi-agent coalitions acting and cooperating in open, unpredictable environments call for interactive algorithms, i.e. algorithms interacting with the environment during performing particular steps of computation or chang-

 ISBN: 978-1-61208-273-8

ing themselves during the process of computation. All of these challenges are present in a domain of traffic control and modeling and can be approached using Perception Based Computing paradigm [25], [26], [27].

Coordination and control are essentially perception based. We understand perception as a process of interpreting sensory data. In the case of road traffic, sensory data can be acquired from traffic control systems as well as from traffic simulators. A crucial issue is how to apply such lower-level data to reason about satisfiability of complex vague concepts including complex spatio-temporal concepts as the concept of a traffic jam or traffic congestion leading to a traffic jam. Complex vague concepts can be used as guards for actions or invariants to be preserved by agents. Such reasoning is often referred as adaptive judgment [10]. Vague concepts can be approximated on the basis of sensory attributes rather than defined precisely. Approximations usually need to be induced by using hierarchical modeling. Unfortunately, discovery of structures for hierarchical modeling is still a challenge. On the other hand, it is often possible to acquire or approximate them from domain knowledge. Given appropriate hierarchical structures, it becomes feasible to perform adaptive judgment [10], starting from sensory measurements and ending with conclusions about satisfiability degrees of vague target guards.

III. Vehicular traffic as a complex system

Vehicular traffic in cities may be considered as a complex dynamic system, which consists of hundreds of thousands independent agents (cars), which drive in the road network realizing a specific goal. This goal is usually reaching a destination point located somewhere in the road network, fulfilling some additional conditions, e.g. minimizing travel time, fuel consumption etc., and following the rules of drive. Agents interact with each other since they use the same road network. This interaction may be purpose of exhibiting new properties of the traffic, such as formation of traffic jams. This property is not obvious from the properties of individual agents (cars) and it is very difficult to predict this phenomenon in advance (e.g. $5 - 10$ minutes before jamming) and to prevent it. In order to ensure collision-free drive of cars, traffic engineers introduce mechanisms, e.g. traffic control systems such as traffic signals at crossroads, which control drive of cars and optimize the traffic.

Despite many years of extensive research, it is still difficult and challenging task to model the traffic in cities properly and with satisfactory accuracy using standard mathematical tools or computer simulations. In addition, the phenomenon may be even more complex and difficult if we assume, that drivers know the real state of the current traffic and choose their routes adaptively. Similarly, it is possible that the traffic control system adapts to the traffic in order to optimize it, which makes the traffic prediction and modeling even more complex. In this kind of complex processes, often the only possible way to model and analyze the process is by making a computer simulation.

IV. Modeling and simulating vehicular traffic

A. Early models

From few decades scientists and traffic engineers have been working on modeling and better understanding the vehicular traffic. They created complex mathematical models, often based on analogies to other real physical phenomenons. For example, some interesting results were obtained by investigating analogy of the vehicular traffic to fluid dynamics. However, traffic flow is significantly different phenomenon, it consists of several substreams, cars have their own start and destination points [12], [23]. There were also approaches to model the traffic using analogies to the kinetic gas theory [20]. These macroscopic models were not able to model the real traffic with satisfactory accuracy. The reason was that they did not take into account local interactions between agents (cars), which are crucial properties of the road traffic.

One of the considered approaches to solve the problem was introducing microscopic models, in which agents (cars) and their interactions were described by mathematical equations, for example *Car-following models* were based on analogy to Newton dynamics equations [23].

B. Models based on cellular automatons

An important progress in modeling vehicular traffic in cities was made by introducing traffic simulation models based on probabilistic cellular automatons. An example of such model is a Nagel-Schreckenberg model (Na-Sch model) [13], [24], which emulates a freeway traffic. Space, time and velocities in the model are discrete, the road is divided into cells, which may be empty or occupied by at least one car, cars motion is defined by properly selected rules. The model was broadly investigated and generalized, e.g. to simulate 2-lane traffic ([21], [14]) or simple crossroads [4].

C. TSF model and software

The Na-Sch model was also used as a base model for a new traffic simulation model developed by the first author of the paper (P. Gora, [6], [8]). The model extends the standard Na-Sch model and enables conducting simulations on a realistic road network, structuralized as a directed graph. The model takes into account, e.g. driver's profile, road's profile, traffic signals, distributions of start and destination points.

The TSF model was later implemented in an advanced software for simulating vehicular traffic in cities, Traffic Simulation Framework. The main window of the software is presented in the Figure 1.

The software uses maps taken from the OpenStreetMap project ([15]) and currently it is able to simulate the traffic in Warsaw. It was confirmed by Warsaw citizens that the

 ISBN: 978-1-61208-273-8

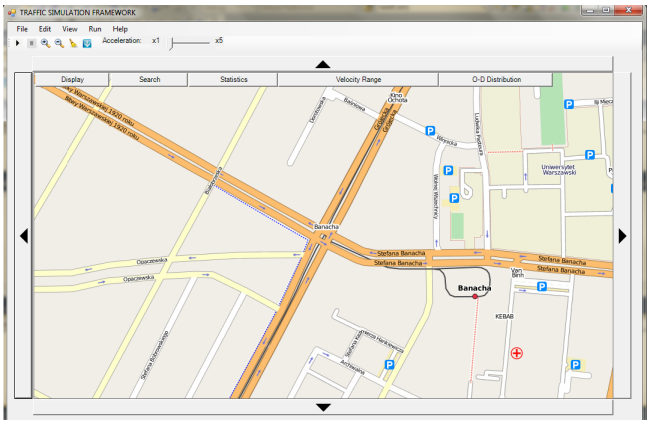

Figure 1. Traffic Simulation Framework - the main window of the application

software can reproduce traffic jams in the same places as they occur in reality. The software is still being developed, its functionality was described in papers [6], [8]. TSF has been already used, e.g. for generating data for the IEEE ICDM 2010 contest [9], [28] and is used by scientists from few countries. The TSF software will be also used for simulating the traffic and interaction with experts. It possesses a multifunctional Graphical User Interface, which can be used to present simulations to experts and acquiring their knowledge about the simulated traffic.

V. METHOD

Traffic is a very complex phenomenon and many high level concepts related to that phenomenon are complex and vague and we do not know how to define it mathematically (also it may depend on many factors such as a type of crossroad, city etc.). However, human brain can recognize such concepts much better. Experts who often drive by cars and stay in traffic jams are able to recognize the traffic situation easily by observing the traffic for some period (e.g. 10 minutes).

In this research, we will acquire such domain knowledge by interaction with experts. We will show short (2-minutes long) movies, presenting simulated traffic, to domain experts. After watching the movie, experts will have to decide what was the state of the traffic. We will also have low level data related to presented traffic situations. Based on the expert domain knowledge and these sensory data, it will be possible to construct hierarchical classifiers (e.g. using rough sets methods [2]) that will approximate the concept of a traffic jam on a given crossroad and extract traffic jams formation models.

Hierarchical classifiers are examples of classifiers which are decision algorithms that map objects to decisions [1], [3]. Objects could be described by low-level numerical or symbolical attributes. Decisions, in many cases, are vague, complex concepts, which are semantically distant from original low-level data. Hierarchical classifiers could be viewed as tools to cover that distance by approximating complex, vague concepts, using low-level data. In such classifiers the classification process goes from input data to decisions through at least few hierarchy levels, from lower data levels to higher, more abstract, complex concept levels. Objects and/or attributes on higher levels are constructed based on objects and/or attributes from lower levels [26], [27]. This process may be supported by domain knowledge, given e.g. in the form of ontologies. To cover the semantic distance, training sets can be constructed with experts support. Decisions could be also complex, temporal or spatio-temporal objects as automated planning of complex objects behaviour, e.g. safe diriving through a crossroad or medical diagnosis, see [2].

In case of our hierarchical classifier, approximating the traffic congestion concept, low-level data, such as number of cars, car's position, current car's speed, will be taken from our traffic simulator ([6], [8]), and decisions will be taken from experts by mean of a dialog. In our traffic congestion hierarchical classifier, objects and attributes from consecutive hierarchy levels will be constructed by information systems, decision tables and decision rules taken from the rough set theory [16], [17], [18] as it was done in [2].

This section describes construction of a traffic congestion training set.

A. Dialog procedure

In this research, we focus on a single crossroad in order to obtain domain knowledge about the traffic congestion and presence of a traffic jam near that crossroad during a given traffic situation, which corresponds to 10 minutes of simulation. This will be done by a dialog with domain experts.

We will conduct some number of simulations using the Traffic Simulation Framework and we will refer to them as *traffic situations* or simply *situations*. For the purpose of the paper we assume to conduct 51 simulations (the proper number should be also subject of further research and experiments). Every situation will last 10 minutes and be run with different parameters, such as:

1) number of cars,
2) start and destination points distributions,
3) initial configuration of traffic lights on a given crossroad.

We selected values of all important simulation parameters in our past research and experiments. These values are presented in the section V-B.

Every situation will be "recorded" - Traffic Simulation Framework will log information about positions and velocities of cars during the simulation in order to read it later and show the same traffic situation to experts using Graphical

 ISBN: 978-1-61208-273-8

User Interface of our software. The following information will be logged to the output file:

- Timestamp (simulation step),
- Car positions (link in the road network, position within the link, geographical longitude and latitude),
- Current car's speed (in km/h).

This logged information enabled reconstruction of the situation, which will be shown as a movie to experts.

We assume that duration of a single phase of traffic lights is constant and lasts 2 minutes for every traffic signal, so every 10-minutes long situation will consist of 5 parts, each of which will last 2 minutes and will correspond to 1 phase of a traffic light. To these situation parts we will refer simply as *phases*.

We divided logs from our 10-minutes long simulations, so it will be possible to show to domain experts 2-minutes long phases separately. Totally, it will give us 255 phases, which lasts 2 minutes each.

Each of 51 situations will be evaluated by domain experts and their task will be to provide information about a traffic state in the area close to the crossroad during every 2-minutes long part of the simulation. In our case (vehicular traffic in cities) a domain expert may be any person who has experience with the city traffic, the most preferable should be drivers, which use road networks in Warsaw often and have to cope with traffic jams.

1 of 51 situations will be analyzed by all experts, while every situation from the rest 50 will be analyzed by 3 experts giving 50×3 situation evaluations. Every expert will analyze 1 situations: 1 common to all experts and 2 taken from the rest 50. Therefore, we will construct 150 / 2 = 75 different tests, one for each expert, so we will need 75 experts. In every test each 10-minutes long situation will be divided into 2-minutes long phases. Thus, every test will be constructed from 15 phases (2-minutes long movies). Additionally, from every situation two phases will be randomly selected to be presented and labeled by experts twice. Therefore, every test will consist of 21 phases presented to the expert in a random order. Experts will not be informed that some of this phases are repeated in a test. After a presentation of a particular movie, the question will be displayed: *What was the traffic congestion?*, and experts will answer the question with one of five possible answers: *Small*, *Medium*, *Large*, *Traffic jam*, *I don't know*. The answer will be provided using the window presented in the Figure 2 which will be shown after every movie.

In the next step, the system will ask experts for the response justification, which they can provide in natural language using the window presented in the Figure 3.

If the user selects *I don't know* response in the first window, the system will ask for checking two closest options from other options available in the window which is presented in the Figure 4.

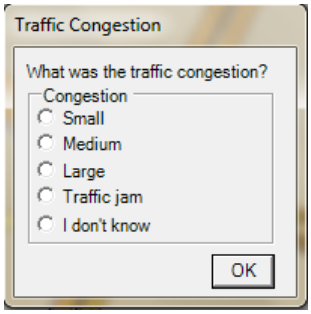

Figure 2. Window that will be shown to experts after every movie

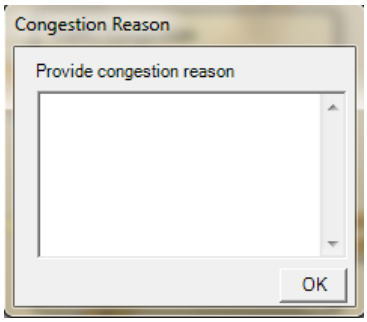

Figure 3. Window to justify the response

After checking the answers and submitting justifications, the next movie will be presented to the expert.

B. Conducting experiments

In our research we will examine the area close to the intersection of streets Banacha, Grójecka, Bitwy Warszawskiej 1920 in Warsaw, which are very close to our Faculty and this crossroad is a place were large traffic congestion occurs very often. The area under investigation is presented in the Figure 5.

We prepared 51 traffic simulation scenarios, each of which will be run using the TSF software producing 51 traffic situations. Every situation lasts 10 minutes and will be run using simulation parameters presented in the table V-B. These parameters were selected based on our preliminary experiments.

Simulations differ in distributions of start and destination points of cars (and routes calculated based on that distributions) and number of cars that start drive every TimeGap steps (V-B). We prepared 5 different distributions of starting points and 5 different distributions of destination points. Distributions of starting points were named "From East", "From West", "From North", "From South", "Uniform", distributions of destination points were named "To East", "To West", "To North", "To South", "Uniform". It gives us 25 configurations of pairs: (start points distribution, destination points distribution). Names of distributions indicates were is the major concentration of start or destination points, respectively. The detailed description of these distributions

 ISBN: 978-1-61208-273-8

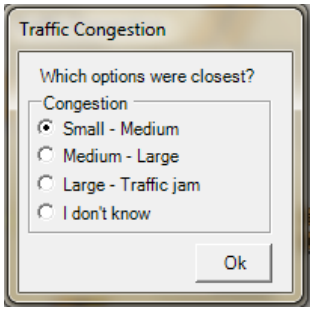

Figure 4. Window for submitting two closest options

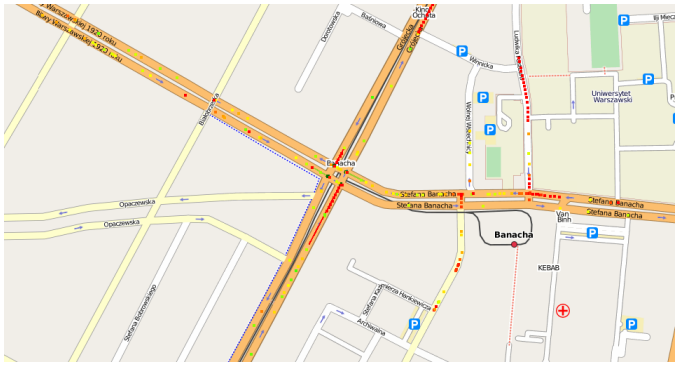

Figure 5. Crossroad of streets Banacha, Grójecka, Bitwy Warszawskiej

and procedures for editing start points and destination points is described in the paper [6].

For every combination of pairs (start points distribution, destination pionts distribution) we still have few degrees of freedom that can be manipulated in order to produce different simulation scenarios. Some of these degrees of freedom correspond to parameters named in the first column of the table V-B: $NrOfCars$, $NewCars$, $Acceleration$, $CrossroadPenalty$, $TurningPenalty$. Other parameters may be related to the initial configuration of traffic signals at the crossroad or maximal velocity permissible on a given street. For our current research we need only 51 simulation scenarios, so we decided to manipulate the parameter $NewCars$. 5 different start points distributions, 5 different destination points distributions and 3 different values of the $NewCars$ parameter gives us 125 possible simulation scenarios, from which we chose 51 that are the most realistic appropriate to conduct our research.

C. Evaluation of obtained decisions

Evaluation of expert decisions can be either *expert-oriented* or *case-oriented*. In the expert-oriented evaluation we will check a consistency of decisions made by a given expert. In this case, the evaluated situation should be labeled by an expert (before evaluation) at least twice for checking stability of the expert decision making. In order to do that, from every situation two phases will be selected to be labeled

Table I
SIMULATION PARAMETERS USED IN OUR EXPERIMENTS

Name of the parameter	Description	Value
NrOfCars	Initial number of cars for a single traffic situation	1000
TimeGap	Time after which new cars start their movement	1 second
Step	Time of a single simulation step	1000 miliseconds
NewCars	Number of cars which start movement after every TimeGap seconds	$5, 3, 1$
Steps	Duration of the simulation	600 simulation steps
Acceleration	Acceleration of cars per simulation step	10 km/h
CrossroadPenalty	Percentage of velocity reduction before the crossroad	25%
TurningPenalty	Percentage of velocity reduction during turning	50%

by an expert twice. In the case-oriented evaluation we will analyze how a given case (phase or situation) is labeled by different experts. For this purpose, every phase will be labeled by three different experts. Their decisions will be used either to determine the final aggregated decision, e.g. by voting, or to find a uniformity of decisions about a given phase. It should be noted that our approach is only one of possible and that decision evaluation itself is a novel and interesting issue and a topic for further research.

VI. CONCLUSIONS AND FUTURE WORK

In the paper, we presented a method for obtaining vehicular traffic domain knowledge using interaction with experts. The method also requires realistic simulations of vehicular traffic, which can be performed using an advanced software Traffic Simulation Framework [6], [7], [8], developed by the first author of the paper. This is still work in progress and presented method will be a subject of our future research. We still need to conduct required experiments and evaluate obtained knowledge. In the next step, we will construct hierarchical classifiers for approximating spatio-temporal complex vague concepts related to vehicular traffic. According to the paradigm of Perception Based Computing, satisfiability of such concepts may activate complex actions, such as reconfiguring traffic lights at crossroads in order to prevent traffic jams or to optimize some key parameters of the traffic. Such classifiers may be used, e.g. for discovering models of complex processes, such as formation of traffic jams, which may be later used to analyze many properties of the traffic. All of this may be a subject of extensive research and obtaining domain knowledge from experts is just the first step. As we argued in the introduction, this step is crucial to construct efficient hierarchical classifiers from low level sensory data in case of such complex process as vehicular traffic in cities.

 ISBN: 978-1-61208-273-8

ACKNOWLEDGMENT

We would like to express our gratitude to Professor Andrzej Skowron for his valuable comments and advices. The research has been supported by the research project realized within Homing Plus programme of Foundation for Polish Science, co-financed from European Union, Regional Development Fund, grant 2011-3/12, and by the grant 2011/01/D/ST6/06981 from the Polish National Science Centre.

REFERENCES

[1] E. Alpaydin, Introduction to machine learning, MIT Press, 2004.

[2] J. Bazan, "Hierarchical classifiers for complex spatio-temporal concepts". Transactions on Rough Sets IX, Lecture Notes in Computer Science, 5390, 2008, pp. 474-750.

[3] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[4] D.Chowdhury and A.Schadschneider, "Self-organization of traffic jams in cities: effects of stochastic dynamics and signal periods", 1999, Physical Review E (Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics), 59(2).

[5] W. Dong and A. Pentland, "A network analysis of road traffic with vehicle tracking data", AAAI Spring Symposium: Human Behavior Modeling, 2009, pp. 7-12.

[6] P. Gora, "Traffic Simulation Framework - a cellular automaton-based tool for simulating and investigating real city traffic", Recent Advances in Intelligent Information Systems, 2009, pp. 641-653.

[7] P. Gora, "A genetic algorithm approach to optimization of vehicular traffic in cities by means of configuring traffic lights", Emergent Intelligent Technologies in the Industry, 2011, pp. 1-10.

[8] P. Gora, "Traffic Simulation Framework", 14th International Conference on Modelling and Simulation, 2012, pp. 345–349.

[9] P. Gora, S. H. Nguyen, M. Szczuka, J. Świetlicka, M. Wojnarski, and D. Zeinalipour, "IEEE ICDM 2010 Contest Tom-Tom Traffic Prediction for Intelligent GPS Navigation", IEEE ICDM 2010 Workshops, 2011, pp. 1372–1376.

[10] A. Jankowski and A. Skowron, "Wisdom technology: A rough-granular approach", in M. Marciniak, A. Mykowiecka (eds.) Bolc Festschrift, Lectures Notes in Computer Science, 5070, Springer Verlag, 2009, pp. 3–41.

[11] R. Keefe, Theories of Vagueness, Cambridge University Press, 2000.

[12] M.J. Lighthill and G.B. Whitham, "On kinematic waves. I. Flood movement in long rivers", Proceedings of the Royal Society of London, Piccadilly, London, A229(1178), 1955, pp. 281-316.

[13] K. Nagel and M. Schreckenberg, "A cellular automaton model for freeway traffic", Journal de Physique, 1992, pp. 2221–2229.

[14] K. Nagel, D. E. Wolf, P. Wagner, and P. Simon, "Two-lane traffic rules for cellular automata: A systematic approach", Physical Review E (Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics), 58(2), August 1998, pp. 1425–1437.

[15] OpenStreetMap, http://www.openstreetmap.org, last accessed: March, 2013.

[16] Z. Pawlak, Rough sets: Theoretical aspects of reasoning about data, Dordrecht: Kluwer, 1991.

[17] Z. Pawlak, "Rough sets", International Journal of Computing and Information Sciences, 18, 1982, pp.341–356.

[18] Z. Pawlak, A. Skowron, "Rudiments of rough sets", Information Sciences 177, 3–27, 2007.

[19] T. Poggio and S. Smale, "The mathematics of learning: Dealing with data", Notices of the AMS, 50(5), 2003, pp. 537–544.

[20] I. Prigogine and R. Herman, Kinetic Theory of Vehicular Traffic, Elsevier, Amsterdam, 1971.

[21] M. Rickert, K. Nagel, M. Schreckenberg, and A. Latour, "Two Lane Traffic Simulations using Cellular Automata", Physica A: Statistical Mechanics and its Applications, 1996, 231(4), pp. 534–550.

[22] J. Stepaniuk, Rough-Granular Computing in Knowledge Discovery and Data Mining, Springer Verlag, 2008.

[23] A. Schadschneider, "Statistical Physics of Traffic Flow", Physica A285, 2000, pp. 101–120.

[24] A. Schadschneider, "The Nagel-Schreckenberg model revisited", The European Physical Journal B, 10(3), 1999, pp. 573–582.

[25] A. Skowron and P. Wasilewski, "An introduction to perception based computing", Lecture Notes in Computer Science, Volume 6485, Springer Verlag, 2010, pp. 12-25.

[26] A. Skowron and P. Wasilewski, "Information Systems in Modeling Interactive Computations on Granules", Theoretical Computer Science, 412, 2011, pp. 5939-5959.

[27] A. Skowron and P. Wasilewski, "Interactive information systems: Toward perception based computing", Theoretical Computer Science, 454, 2012, pp. 240-260.

[28] TunedIT, http://tunedit.org/challenge/IEEE-ICDM-2010, last accessed: March, 2013.

[29] V. Vapnik, "Learning Has Just Started" an interview with Prof. Vladimir Vapnik by R. Gilad-Bachrach, http://learningtheory.org, 2008, last accessed: 2012.

 ISBN: 978-1-61208-273-8