2010 IEEE International Conference on Data Mining Workshops

IEEE ICDM 2010 Contest
TomTom Traffic Prediction for Intelligent GPS Navigation

Marcin Wojnarski*, Paweł Gora†, Marcin Szczuka†, Hung Son Nguyen†, Joanna Swietlicka* and Demetris Zeinalipour‡

*TunedIT Solutions, Zwirki i Wigury 93/3049, 02-089 Warsaw, Poland
Email: {marcin.wojnarski,j.swietlicka}@tunedit.org

† Faculty of Mathematics, Informatics and Mechanics, The University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
Email: {pawelg,szczuka,son}@mimuw.edu.pl

‡ Department of Computer Science, University of Cyprus, 1678 Nicosia, Cyprus
Email: dzeina@cs.ucy.ac.cy

***Abstract*—In this foreword, we summarize the IEEE ICDM 2010 Contest: "TomTom Traffic Prediction for Intelligent GPS Navigation". The challenge was held between Jun 22, 2010 and Sep 7, 2010 as an interactive on-line competition, using the TunedIT platform (http://tunedit.org). We present the scope of the ICDM contest series in general, the scope of this year's contest, description of its tasks, statistics about participation, details about the TunedIT platform and the Traffic Simulation Framework. A detailed description of winning solutions is part of this proceeding series.**

***Keywords*-Contest, TomTom, TunedIT, traffic, simulation, prediction, data mining, Warsaw**

I. The ICDM Data Mining (DM) Contest

The ICDM Data Mining Contest offers a unique opportunity to scientists and enterprises, to involve teams of domain experts that will compete against each other in order to develop and test data mining techniques that can improve real or realistic applications.

The general philosophy of the contest is to provide to participants a set of custom datasets, evaluation metrics (or software tools) as well as expected answers to a set of predetermined tasks. The participants are then asked to identify the best possible solutions to the given tasks maximizing the given evaluation metrics. Competing team work off-line to implement the tasks outlined by the contest organizers. The results of each team are then submitted to the organizers prior the conference date. The contest organizers select the submissions that will be included in the proceedings of the conference. The awarding process is carried out during the conference.

The previous ICDM DM contests were organized as follows: In 2007, the Hong Kong University of Science and Technology (Hong Kong) organized a competition in Omaha, NE (USA) with title: *"Estimating Location Using Wi-Fi"* (http://www.cse.ust.hk/~qyang/ICDMDMC07/). In 2008, the University of Ottawa (Canada), the Universiteit Utrecht (Netherlands) and Health Canada (Canada) organized a competition in Pisa, Italy with title: *"Radioxenon monitoring for verification of the Comprehensive nuclear-Test-Ban Treaty"* (http://www.cs.uu.nl/groups/ADA/icdm08cup/). In 2009, the Walt Schneider Research Lab at the University of Pittsburgh (USA) organized a competition in Miami, FL USA with title: *"The Pittsburgh Brain Competition"* (http://pbc.lrdc.pitt.edu/). This year, TunedIT Solutions (Poland), the University of Warsaw (Poland) and TOMTOM International BV (The Netherlands), organize a competition in Sydney Australia with title: *"Traffic Prediction for Intelligent GPS Navigation"* (http://tunedit.org/challenge/IEEE-ICDM-2010)

II. The 2010 Contest Description

Over the last century the number of cars engaged in vehicular traffic in cities has increased rapidly, causing many difficulties for all citizens: traffic jams, large and unpredictable communication delays, pollution etc. Excessive traffic became a civilization problem that affects everyone who lives in a city of 50,000 or larger, anywhere in the world. Complexity of processes that stand behind traffic flow is so large, that only data mining algorithms may bring efficient solutions. With the IEEE ICDM 2010 Contest, the organizers asked researchers to devise the best possible algorithms that tackle problems of traffic flow prediction, for the purpose of intelligent driver navigation and improved city planning.

The challenge was organized in the form of an interactive on-line competition, at TunedIT platform, between June 22, 2010 and September 7, 2010. The winners were awarded prizes worth $5000, sponsored by TomTom (http://www.tomtom.com/).

Organizing Committee of the challenge consisted of four members: Marcin Wojnarski, Paweł Gora, Hung Son Nguyen and Marcin Szczuka.

The contest was sub-divided into three independent tasks:

1) **Task 1** – Traffic congestion prediction, in an elementary setup of time series forecasting: a series of measurements from 10 selected road segments is given

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.51

and the goal is to make short-term predictions of future values based on historical ones.

2) **Task 2** – Modeling the process of traffic jams formation during the morning peak in the presence of roadwork, based on initial information about jams broadcast by radio stations. Input data contain identifiers of road segments closed due to roadwork, accompanied by a sequence of segments where the first jams occurred. The algorithms had to predict a sequence of segments where next jams occur in the nearest future.
3) **Task 3** – Traffic reconstruction and prediction based on real-time information from individual drivers. Input data consists of a stream of notifications from 1% of vehicles about their current locations in the city road network, sent every 10 seconds. The algorithm received this stream and had to predict the traffic congestion on selected road segments for the next 30 minutes. Large volumes of data were involved in this task, requiring the use of scalable data mining methods.

As tasks were independent, anyone could participate in all of them or in a chosen one.

All three tasks address the same problem of traffic congestion prediction, but each of them approaches this problem from a different perspective: employs data of different characteristics and structure, expects different output decisions from the algorithm, has different levels of difficulty and requires the use of different data mining techniques. A graph topology of streets was available to participants. To find the most effective algorithms, they had to mine these structured data and exploit regularities observed for different nodes, segments and subgraphs.

All the tasks are very important for practical applications. They are also related to the most current and hotly debated topics in data mining research: mining structured data, mining data streams, large-scale and temporal data mining.

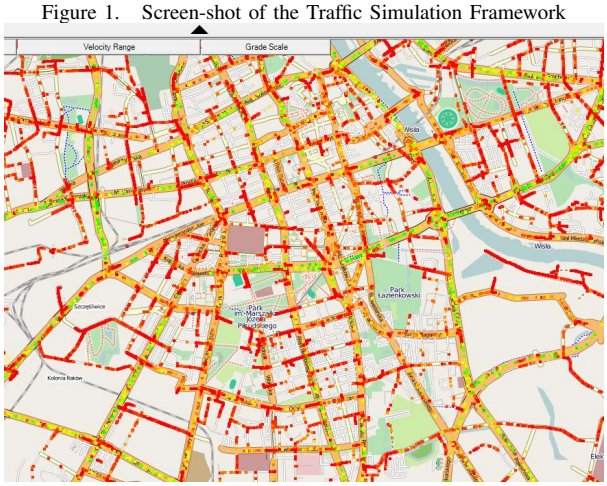
III. Traffic Simulation Framework

Competition datasets come from a simulator of vehicular traffic, the Traffic Simulation Framework (TSF) [1], being developed at the University of Warsaw since 2007. Simulations employed authentic map of the city of Warsaw, Poland, taken from OpenStreetMap project.

Thanks to the use of TSF, we had knowledge about traffic on any given edge of the street graph, at any point in time, so we could provide data that would have been unavailable otherwise, like target decisions in the 3rd problem. We could also formulate more challenging and interesting tasks.

Simulations in TSF are based on the most well-known, realistic model of Nagel and Schreckenberg [2], used also in the German project of Autobahn Traffic in North Rhine-Westphalia [5]. The model describes traffic on straight, one-way road, divided into some number of cells. Each cell may contain one car or may be empty. In each tick of a discrete clock, cars change their positions and speeds with carefully chosen rules [1].

Figure 1. Screen-shot of the Traffic Simulation Framework



Nagel-Schreckenberg (NS) model was broadly examined and generalized [3], [4]. In particular, it was shown that this model can reproduce the phenomenon of traffic jams with unprecedented accuracy. TSF extends the NS model to simulate traffic on an arbitrarily complex graph of road segments and incorporates new rules to model crossroads, traffic lights, multi-lane streets etc. Each car is an autonomous agent, with its own start and destination points, as well as preferences regarding its road behavior. Road segments may differ in type (minor/major), default maximal velocity and number of lanes - authentic data are included in maps from OpenStreetMap. TSF enables users to edit configuration of traffic lights, distribution of start and destination points, and many other simulation parameters. It was confirmed by Warsaw citizens who are intensive car drivers that TSF very accurately reproduces traffic jams, in the same places where they occur in reality.

After the contest, TSF will be released for public use, to enable post-competition research.

IV. Results

The contest attracted 575 participants (both teams and individuals), of whom over 100 submitted solutions, most of them several times: the total number of solutions was nearly 5000. Best algorithms achieved nearly 3-fold improvement over baseline solutions in predicting traffic congestion and jams.

The results obtained by the three best participants in each track, together with baseline results, are presented in tables I, II, and III, respectively. The symbol (*) next to the participant's rank denotes that there is a report describing their solution in this proceedings.

All resources, data, TSF and information from competition discussion blog and discussion list are now publicly

Table I
FINAL RESULTS OF THE TRAFFIC TASK

Rank	Participant or Team	Username	Result
1	Alexander Groznetsky, Ukraine	alegro	25.2327
2*	Carlos J. Gil Bellosta, *Datanalytics*, Spain	datanalytics	25.4167
3	Benjamin Hamner, *Duke University*, USA	hamner	25.4337
—	Baseline	—	44.9444

Table II
FINAL RESULTS OF THE JAMS TASK

Rank	Participant or Team	Username	Result
1*	Lukasz Romaszko, *University of Warsaw,* Poland	lukasz212121	0.56056
2*	Jingrui He, Qing He, Grzegorz Swirszcz, Yiannis Kamarianakis, Rick Lawrence, Wei Shen and Laura Wynter, *IBM T.J. Watson Research Center*, USA	trafficlab	0.55793
3	Kenneth Shirley, Carlos Scheidegger, Ji Meng Loh and Suhris Balakrishnan, *AT&T Labs Research*, USA	regress	0.55614
—	Baseline	—	0.42805

available. Everyone can use them as a start point for new research.

V. TASKS

A. Task 1 - "Traffic": Traffic congestion prediction

Many municipalities use devices called Automatic Traffic Recorders (ATR) to collect traffic data from a number of selected road segments in the city. ATRs are magnetic loops embedded in the pavement surface, detecting the presence of metal and transforming this information to volume data.

Table III
FINAL RESULTS OF THE GPS TASK

Rank	Participant or Team	Username	Result
1*	Benjamin Hamner, *Duke University*, USA	hamner	6.7719
2*	Jingrui He, Qing He, Grzegorz Swirszcz, Yiannis Kamarianakis, Rick Lawrence, Wei Shen and Laura Wynter, *IBM T.J. Watson Research Center*, USA	trafficlab	7.4556
3	Andrzej Janusz, *University of Warsaw,* Poland	NosferatoCorp	7.5779
—	Baseline	—	18.0649

Data from ATRs are used further on to make short-term and long-term predictions, for the purpose of driver navigation and urban planning (roadwork, new road construction etc.), as for example in the Autobahn Traffic project [5].

In this task, we asked participants to devise algorithms for making predictions of this kind. We simulated a time series of congestion measurements produced by ATRs installed on 10 selected road segments in the city of Warsaw and asked participants to make short-term predictions of future values based on historical ones. This task was intended as an introductory one. It used an elementary framework of time series forecasting, so that everybody could try to solve it. Despite the simplicity of the problem formulation, its solution has practical importance both for driver navigation and city planning.

Data consist of a time series of 20 congestion measurements, 2 values for each of the 10 selected road segments, corresponding to two opposite directions of traffic flow. Congestion – the number of cars that passed a given segment – was measured in consecutive 1-minute periods of time. The TSF simulator worked in 10-hour long simulation cycles. During the cycle, distributions of start and destination points of new vehicles were exchanged every 60 minutes. After 10 hours, simulation was restarted from scratch. Distributions were selected randomly from a pool of 10 predefined ones, manually designed and verified against real-world plausibility. Additionally, small random fluctuations were imposed on selected distributions every time and configuration of traffic lights was modified.

The training dataset, publicly available for participants, consisted of a continuous stream of data collected over 1000 hours of simulation, divided into a hundred of 10-hour long independent cycles. The test dataset covered different 1000 hours, and was split into 60-minute long windows, of which only the first 30 minutes were revealed to participants, while the other 30 minutes were left for evaluation of predictions. The task was to predict congestion (total number of cars) for time period between 10 and 20 minutes ahead, i.e., for the period between 41'st and 50th minute of every window. Windows in the test dataset were permuted, so that participants could not deduce future congestion by looking at the following time window.

The baseline solution was calculated as a total number of cars in the last ten minutes of the known part of the window (minutes 21'st till 30th).

Solutions were evaluated by the Root Mean Squared Error (RMSE) of predictions.

B. Task 2 - "Jams": Modeling the process of traffic jams formation in the presence of roadwork

It was a long time ago when radio stations came up with an idea of collecting current information about traffic jams, roadwork or accidents, and radio broadcasting it to all the drivers, so that they can go around the impassable places.

In this task we wanted to go one step further and try to mine the data gathered during the initial phase of the morning peak, in order to predict where the next set of jams is going to happen during the main phase of the peak. Such predictions could be used to warn drivers in advance, before the jams actually occur.

Data collected by radio broadcasters have some distinct features that influence the way they should be mined and with which algorithms. These features were reflected in the competition data. Firstly, they are imprecise: detailed measurements of speed and number of cars, as well as the time a traffic jam first occurred, are missing. Secondly, they cover only major roads that constitute at most 25% of the whole road network. The missing part of information, related to minor roads, may influence jams on major roads and thus should be included, e.g., as latent variables, in decision models. Competition data exhibited the same properties: it contains ordered sequences of street identifiers alone, without any numeric information on congestion or timing of jams, covering only major streets of Warsaw.

Input data contained identifiers of 5 road segments closed due to roadwork – 2 on major roads and 3 on minor – accompanied by a sequence of segments of major roads where the first jams occurred during the initial 20 minutes of the simulation. The task was to predict a sequence of segments of major roads where next jams occur in the next 40 minutes. The length of both sequences could vary between samples. Graph of Warsaw's streets was given to participants, so that they could exploit the structural dependencies between different parts of the network.

The criterion that defined a jam involved both the average speed and number of cars on a given segment. A segment was said to be jammed if the average speed over the last 6 minutes was lower than 5 km/h and the number of cars that have passed or stayed on this segment was larger than 10. A given road segment could be listed at most once during the simulation, when it got jammed for the first time.

Data samples were generated from independent 1-hour long simulations, starting each time with an empty road network, fed subsequently with cars according to the selected distributions of start and destination points. Selection of distributions was done in the same way as for Task 1. Locations of roadwork were selected randomly. Training and test sets consisted of 5000 samples (hours) each. In the test set, only the input part of samples was revealed, while the output part - the sequence of next predicted jams - was submitted by participants as a solution.

The baseline solution predicted always the most frequently jammed edges that did not appear in the input part of a given sample. The length of the prediction was equal to the average length of output sequences in the training data. Frequency of jams on a given edge was calculated over the whole training data.

Due to atypical form of output decisions - varying-length sequence of identifiers - we employed a non-standard evaluation metric, based on the concept of Mean Average Precision (MAP), from the domain of Information Retrieval, adapted to the specifics of this task. For each sample, quality of prediction was calculated as:

$\frac{1}{N} \cdot \sum_{i=1}^{N} Precision(i)$, where:

$$Precision(i) = \frac{|P_i \cap T_i|}{i},$$

P - is the predicted sequence of identifiers, T - is the target sequence of identifiers, P_i, T_i - are sets of first i elements of P and T, $N = max(length(P), length(T))$.

This measure reaches the largest value when both sequences are exactly equal and punishes any deviations of predictions from the target, like: different length of the sequence, different order of identifiers, lack of expected identifiers or presence of non-expected identifiers. Mistakes on initial positions of the sequence are punished stronger than on further ones.

Overall result for the test set was calculated as an arithmetic average of values calculated for each sample.

C. Task 3 - "GPS": Smart GPS navigator

Recently two independent technologies - GPS car navigation and wireless Internet access in cell phones - became so popular, that many drivers use them both and thus can send their current GPS positions to the server in real time. Stream of such data coming from different drivers can be merged to reconstruct current map of traffic in the whole city and make predictions for the next period of time, using data mining methods. These predictions would be sent back to drivers to be employed in smart real-time journey planning: choosing faster routes and optimizing global traffic in the city.

In the third task we asked participants to devise algorithms for solving this problem. We assumed that 1% ($\sim$500) of drivers in TSF use GPS navigation, which sends a notification to the central server about its current GPS position and velocity every 10 seconds. Algorithms designed by participants received this stream as an input. The stream covered a 30-minute interval, and the participant had to predict average velocity of vehicles passing 100 selected road segments in 6-minute time periods: from now on until 6 minutes ahead, and between 24th and 30th minute ahead. Harmonic mean was used instead of arithmetic mean, as it corresponds better to travel times, which are the ultimate criterion of optimization in a real-world setting. Another requirement for the algorithms was scalability, since the provided sets of data were highly voluminous: the training set had several GBs.

Training and test datasets covered 500 hours of simulation each. The same type of simulation as in the "Traffic" task, consisting of 10-hour long cycles, was used. Test set was split into one-hour windows, half of each window was revealed, and another half was to be predicted by participants.

The baseline solution was the average velocity of all cars that passed through each of the edges in the whole given 30-minute long period. If there was no such car, the overall average was taken.

Solutions were evaluated by the Root Mean Squared Error (RMSE) of inverted predictions. That is, predicted velocities were transformed - through inverting and multiplying by 60 - into predicted travel time over 1 km of the road segment, expressed in minutes. These travel times were compared with ground truth using RMSE.

VI. TunedIT platform

The contest was hosted on the TunedIT platform [6]. TunedIT delivers a set of web-based tools that facilitate experimental investigation and scientific collaboration in the field of data mining and machine learning, especially the automation of experiments and the generation of reproducible experimental results. These tools include also a framework for data mining challenges and competitions (http://tunedit.org/challenges), where different scientific events are periodically organized.

The use of TunedIT Challenges platform has many advantages:

- Participant registration, submission of solutions and publication of results are managed by TunedIT website.
- Live leader-board, automated evaluation, multiple submissions: solutions are evaluated automatically, instantly after their submission, and preliminary results are published on Leader-board at the challenge web page. Participants are allowed to submit solutions many times, for the whole duration of the challenge, so they have opportunity to compare their algorithms with others' and make improvements. This gives great fun for participants and encourages high activity.
- Distinct preliminary and final evaluation: to avoid overfitting to test data used in preliminary evaluation - which is possible in the presence of live leader-board and multiple submissions - there is a separate evaluation performed at the end of the contest, employing a distinct dataset. This guarantees that the final results are non- biased.
- Post-challenge research: scientific challenges at TunedIT have their continuation even after final results are announced. Challenge resources - test datasets, evaluation procedures, participants' solutions etc. - are published on-line in Repository, so new algorithms can be tested against challenge data, using the same experimental setup. The contest contributes to creation of benchmarks that can be used in future research by the whole scientific community.
- Post-challenge submissions: TunedIT provides a versatile framework for evaluation of new solutions, submitted after the end of the contest, against challenge datasets, in the same experimental setup. This framework employs TunedTester application and a database of experimental results.

Acknowledgments

We would like to thank the Conference Chair, Dimitrios Gunopulos (University of Athens, Greece) and Chengqi Zhang (University of Technology, Sydney, Australia) for their frequent advice that guided us through many of the questions and concerns that arose along the way. We also thank the rest organizing committee for taking care of dissemination, the proceedings and the local arrangement.

Our special thanks go to TomTom, the sponsor of the competition, and to Mrs. Hanna Gronkiewicz-Waltz, the Mayor of Warsaw, who was the honorary patron.

This research has been partially supported by grants N N516 368334 and N N516 077837 from the Ministry of Science and Higher Education in the Republic of Poland.

References

[1] P. Gora, *Traffic Simulation Framework - a cellular automaton-based tool for simulating and investigating real city traffic*, Recent Advances in Intelligent Information Systems, p. 642-653, Warsaw, 2009.

[2] K. Nagel, M. Schreckenberg, *A cellular automaton model for freeway traffic*, Journal de Physique, p. 2221-2229, 1992.

[3] D. Chowdhury, A. Schadschneider, *Self-organization of traffic jams in cities: effects of stochastic dynamics and signal periods*, Physical Review E (Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics), Volume 59, Issue 2, p. 1311-1314, 1999.

[4] A. Schadschneider, *The Nagel-Schreckenberg model revisited*, The European Physical Journal B, Volume 10, Issue 3, p. 573-582, 1999.

[5] S. Marinsson, R. Chrobok, A. Pottmeier, J. Wahle, M. Schreckenberg: *Simulation Framework for the Autobahn Traffic in North Rhine-Westphalia*, in International Conference on Cellular Automata for Research and Industry, ACRI, LNCS 2493, p. 315-324, 2002.

[6] M. Wojnarski, S. Stawicki, P. Wojnarowski, *TunedIT.org: System for Automated Evaluation of Algorithms in Repeatable Experiments*, in M. Szczuka et al. (Eds.): Rough Sets and Current Trends in Computing 2010, LNAI 6086, p. 20-29. Springer, Heidelberg, 2010.