orig-research
Intelligent Transportation Systems
Editor: Rosaldo J.F. Rossetti, University of Porto, rossetti@fe.up.pt

The Role of Context in Transport Prediction

Francisco C. Pereira
Singapore-MIT Alliance for Research and Technology

Ana L.C. Bazzan
Universidade Federal do Rio Grande do Sul

Moshe Ben-Akiva
Massachusetts Institute of Technology

After a few decades of research and development in intelligent transportation systems (ITS), we have an impressive amount of hardware and software tools that can monitor, estimate, and control the traffic network. Such tools are essential for traffic management and traveler's decision making, but the complex role of human behavior in the transportation system demands considerations that might not be captured with sensors that are focused on the network or vehicles. For example, the traffic manager needs to understand why certain congestion is formed (Is it an incident? A special event? A religious ceremony? Weather? School pick-up/drop-off?) and to predict how it will evolve. A special event leads to different patterns and management procedures than an incident or a flooding event. In other words, besides knowing that a problem exists, traffic managers and prediction systems need to know its *context*.

So, what's missing? How can we extend current ITS technologies to capture and process such information? Here, we suggest that the Internet is a resource for contextual information and we'll overview available techniques and open questions to use it in ITS, particularly transport prediction.

What's Missing: Basic Parameters and Challenges

Let's start with a definition. We consider *context* as any available semantic information that can be associated to observations from the traffic-sensing system (for example, cameras, loop counters, and GPS probes). This semantic information can come in the form of webpages with natural language text, records in a database, official RSS feeds (on weather, incidents, and road work), microblogs (such as Twitter), social networks, location-based services (such as Waze), or other publicly available resources with relevant data. In fact, we can go beyond the Internet; in theory, we could include radio stations, or even private data (SMS), but these are generally non-accessible.

Context can be important to explain and help predict many transport-related phenomena. For example, a sudden demand peak in an area can be due to special events, religious activities, political demonstrations, street fairs; general demand pattern changes can be associated to school holidays; and non-recurrent supply changes can be caused by incidents, road works, road blockages, and harsh weather. On a somewhat different perspective, context can be used to analyze aspects transversal to behavior and transport, such as well-being (for example, sentiment analysis on public transport) or environment (online reports on emissions).

Knowing context is particularly relevant in non-habitual scenarios. While in recurrent scenarios, traffic managers and commuters are aware of their evolution and available options, in nonrecurrent ones, they need good predictive capability to make decisions. Adding semantic causal information to the prediction process together with observation should contribute to its accuracy.

Interestingly, the Internet is a privileged dissemination means regarding non-habitual circumstances: while people don't often tweet, post, or consult the Web about their usual commute, they're more likely to do so if something abnormal happens; special events organizers will use the Internet to reach their public; many newspapers post their articles online even before the paper edition; and authorities use the Web to post warnings on weather, infrastructure changes, incidents, detours, new bus lines, and so on.

We identify three major challenges with respect to extending current ITS solutions with relevant context. These are the focus of our article:

- Where is the relevant information? This is known as the problem of *information retrieval* (IR), and can be formulated as the task of obtaining the list of documents that best matches a given query. Google search is the best-known example of a general-purpose information retrieval engine.
- Even if we have the best documents, most of them are in unstructured text, and some are structured but

Submission draft version. This article was published in IEEE Intelligent Systems, vol. 29, no. 1, pp. 76-80.

useful information isn't explicit. How can we turn such data into relevant information that's understandable by ITS systems? This is known as an *information extraction* problem.

- How can we use such information for *predicting* the transportation system's state? Particularly for non-habitual scenarios, context attributes might be statistically scarce and thus have limited predictive power.

Information Retrieval

When we're looking for information on the Web, we can consider two resources: specific websites/databases that we know well, or generic Web search engines such as Google or Bing that can retrieve any document in the Web. In both cases, the first challenge is to define what we're looking for, the *query*. An efficient query is as specific as possible to only retrieve relevant documents and as generic as possible to avoid ignoring important ones. To build efficient queries for ITS, we can take advantage of spatial and temporal constraints (for example, congestion happens in a certain area at a certain time) together with the specific scenario in question. This is helpful to constrain the search to certain websites. For example, demand hotspots may be explained with events' websites or Facebook, while incidents may be mentioned in official feeds, Twitter, or news websites.

Constraining search to specific resources helps reduce the IR problem to a trivial one: if we know well the resource (for example, an events announcement website or Twitter), retrieval is reduced to simple customized calls (getting events from location X at time Y, or getting all georeferenced Tweets from an area). Of course, this approach implies manual selection of resources and tailoring of query scripts, and it's constrained to a limited number of scenarios. For each scenario, there are specific ideal databases. For example, for demand prediction for special events scenarios,[1,2] the authors used events websites to automatically obtain data (e.g. Eventful.com). In other scenarios, Twitter has been used to retrieve relevant information[3] about incidents or natural disasters.[4] Other resources can be named, such as Waze, Google trends, or Facebook.

The option of generic search poses complicated challenges in terms of quality. It's well-known that, due to ambiguity in language, queries in the unconstrained Web easily bring a lot of repetition, spam, and wrong pages.[5-7] On the other hand, if a relevant document exists anywhere online, in theory, it's possible to retrieve it. In the ITS context, the generic IR approach is still an open problem as far as we know. From the perspective of a roadmap for research and practice, we argue that constraining IR to specific websites/databases is, at this stage, the practical option. Ultimately, a fully open system is desirable, to cope with every possible scenario. This is an open research opportunity, and plenty of literature exists in IR[8] (including the ACM Special Interest Group on Information Retrieval conference series) to build on.

Information Extraction

Having the right document at the right time isn't sufficient—we need to translate such data into features that are useful for ITS applications. We want to extract information about *what* (for example, an incident, concert, sports game, or religious celebration), *when*, *where*, and other relevant attributes (including how many lanes are blocked in an incident, the cost of concert tickets, the public's age range, or a temple's size). This is generally seen as an information extraction (IE) problem. Sometimes this information can be easily obtained through APIs or screen scraped from well-structured websites (when such practices are allowed). This was the technique used earlier in demand prediction in special events scenarios,[1,2] where the event type, location, and start time were used as input to a neural network model.

Unfortunately, APIs with adequate, well-tailored information won't always be available, particularly for unstructured text. Much more complex IE techniques are needed to extract relevant information, such as keyword/key phrase extraction,[9] named entity recognition,[10] or latent topic modeling.[11] In each case, the extracted information can participate in an ITS prediction model as the dummy variable (presence/absence of keyword) or continuous value (frequency of word or topic).

In their work on incident impact prediction, Mahalia Miller and Chetan Gupta[12] apply a rule-based string-matching technique to extract a limited number of keywords (such as the vehicle type) from real-time incident report messages. A similar approach is followed by Eric Mai and Rob Hranac,[3] applied to Twitter messages, for incident analysis and detection. In other work,[13] we used topic modeling to extract implicit topics that help predict incident clearance time. The latter technique has two important advantages over rule-based keyword extraction approaches: it's fully automated so there's no need to hand code rules for specific keywords; and given its probabilistic foundations, it provides a continuum of incident severity causes (for example, incident *X* has *Y* percent of an injury-related topic, or *Z* percent of an oil-spillage topic). The same technique was applied recently to break down special event public transport demand hotspots into their constituents (for example, a hotspot is caused by *X* percent of people from event *A*, *Y* percent from event *B*).[14]

In a different vein, sentiment analysis also applies IE to extract polarized opinions or emotional content from natural language text. It has been extensively applied in market research and social network analysis. In transportation, sentiment analysis has been applied over Twitter data to assess public transport passenger satisfaction,[15,16] but it's potentially relevant in other types of applications (including the detection of crisis situations, public demonstrations, congestion, and assessing special popularity).

Many other types of contextual data are available for ITS, such as weather information, school calendars, religious ceremonies, or road work information. Due to its heterogeneity, the IE task needs to be considered independently in each case. Fortunately, plenty of open source tools exist that have all the essential algorithms for topic modeling, named entity recognition, or key phrase extraction, such as Mallet[17] and LingPipe[18] in Java, or Gensim in Python.[19] These also include tools such as stopword lists, parsers, stemmers, and noun-phrase chunkers, that are essential for data preparation by removing unnecessary words (such as the coordinating conjunctions "and" and "or"), grouping them, or re-representing them (for example, by their word stem or syntactic role).

The bottom line is that, after retrieving the right documents, the user needs to obtain the features that are relevant for the ITS application at hand. To the best of our knowledge and intuition in ITS, this process should be considered on a case-by-case basis, taking advantage of the available data and tools.

Transport Prediction with Context

Due to its heterogeneity, the first challenge of using context data for prediction is how to represent it. The previous section gave clues on how to do it: unstructured text can be re-represented as fixed sets of topics much in the same way that a continuous signal can be re-represented by principal components analysis (PCA) eigenvectors; keywords can be identified through rules and be represented with dummy variables; and in fact, other attributes may exist with numeric quantities (such as how many lanes are blocked in an incident, or the cost of concert tickets), and space or time attributes.

The next challenge relates to an apparent paradox in scenarios where context is more valuable—for example, non-recurrent conditions such as incidents, strong weather, or special events. How can we predict if these are uncommon? Do we have enough evidence in the database to generalize? Approaches that depend only on a rich historical database might not be adequate: statistical inference or machine learning algorithms alone will try to look for *similar* past cases and find none or miss important details. In fact, even with a lot of historical data from related scenarios, subtleties may undermine the prediction. For example, incidents may lead to very different patterns than apparently similar ones in the past, depending on how much capacity is reduced, its precise location (such as before or after off-ramp, and visibility), destinations of affected traffic, and network topology (for example, the affected link's centrality). Even with highly accurate contextual information,[3,12,13] there's a strong chance that similar past cases will yield totally different outcomes.

To solve this problem, we need to simulate the transportation system considering travelers' decisions. This is the proposal of real-time dynamic traffic assignment (DTA) models such as DynaMIT[20] or DynaSMART,[21] which are able to efficiently simulate traffic flows taking into account individual driver's choices. To make a prediction in a real-time context, first they need to self-calibrate the supply and demand parameters through an optimization process guided by sensor observations (such as loop counts, speeds, travel times, and GPS data). The ultimate goal of this online calibration process is to understand where people are going (and in which mode), and the network performance at the link level (for example, what are current capacities).

Context data can help a DTA-based prediction at two points: providing initial parameters for calibration (such as capacity reduction in an incident area,[13] or expected speeds/densities during rain);[22] and predicted parameter changes (for example, when an incident will be cleared,[13] or when people will go to a special event).[2] The role of the DTA is to put all these together in a simulation that considers human behavior such as mode, route, or departure time choices.[23]

Statistical inference and machine-learning algorithms can provide crucial help in this process. Due to real-time constraints and search space complexity, online DTA calibration heavily depends on the starting, a priori, parameters. Many such algorithms are very efficient in getting a good first guess. For example, neural networks, radial basis functions, support vector machines, hazard-based functions, graphical models, and many others have been used successfully for incident analysis,[13,24,25] special events demand prediction,[1,2,14] or speed/density parameters inference.[26]

Thus, using IR and IE techniques, we can capture and extract contextual information for ITS applications. Using machine-learning methods, we can translate such data into transport-relevant parameters such as capacity reduction or demand flows. We argue that, particularly in non-recurrent scenarios, we need to apply transport prediction methods that rely on travel behavior models, such as DTA simulators.

Looking at the bigger picture, the next generation of intelligent transportation systems needs to be context-aware, and able to consider multiple explanatory sources that exist beyond network monitoring technologies. Cisco has forecasted a three-fold growth in global Internet traffic from 2012–2017,[27] and the smartphones market is expected to keep growing at a high pace (32 percent in 2013), particularly in the developing world.[28] This will affirm the Internet as a central resource about events in the city, from small incidents to major situations such as large concerts or natural disasters. From our perspective, the opportunities and open scientific challenges are immense and generally still at an early stage.

Context mining is complementary to network-sensing technologies (see Figure 1). While the latter provides information on *what* is happening in traffic, the former helps understand *why*. When properly aligned in space and time, they become essential to understand how traffic will evolve, by being inputs to transport-prediction algorithms. We particularly argue that, with the transportation system ultimately being driven by human behavior, such predictions need to take into account the traveler's choices with respect to perceived context and traffic conditions.

Figure 1. The bigger picture. Context mining is complementary to network-sensing technologies; and while the latter provides information on what is happening in traffic, the former helps understand why. When properly aligned in space and time, they become essential to understand how traffic will evolve, by being inputs to transport-prediction algorithms.

References

1. F. Calabrese, F. C. Pereira, G. D. Lorenzo, L. Liu, and C. Ratti, "The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events," *Pervasive Computing*, LNCS 6030, Springer, 2010, pp. 22–37.
2. F.C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Using Data from the Web to Predict Public Transport Arrivals under Special Events Scenarios," *J. Intelligent Transportation Systems*, preprint, 2013.
3. E. Mai and R. Hranac, "Twitter Interactions as a Data Source for Transportation Incidents," *Proc. Transportation Research Board 92nd Ann. Meeting*, 2013, no. 13-1636; http://docs.trb.org/prp/13-1636.pdf.
4. A. Crooks et al., "#Earthquake: Twitter as a Distributed Sensor System," *Trans. GIS*, vol. 17, no. 1, 2012, pp. 124–147.
5. Y. Matsuo, H. Tomobe, and T. Nishimura, "Robust Estimation of Google Counts for Social Network Extraction," *Proc. Nat'l Conf. Artificial Intelligence*, AAAI Press, 2007; www.aaai.org/Papers/AAAI/2007/AAAI07-221.pdf.
6. R. Wicklin, "Estimating Popularity Based on Google Searches: Why It's a Bad Idea," *The Do Loop*, blog, 19 Aug. 2011; http://blogs.sas.com/content/iml/2011/08/19/estimating-popularity-based-on-google-searches-why-its-a-bad-idea.
7. A.O. Alves et al., "Place in Perspective: Extracting Online Information about Points of Interest," *Ambient Intelligence*, LNCS 6439, Springer, 2010, pp. 61–72.
8. C.D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, 1st ed., Cambridge Univ. Press, 2008.
9. S.N. Kim et al., "Semeval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles," *Proc. 5th Int'l Workshop on Semantic Evaluation*, Assoc. Computational Linguistics (ACL), 2010, pp. 21–26.
10. L. Ratinov and D. Roth, "Design Challenges and Misconceptions in Named Entity Recognition," *Proc. 13th Conf. Computational Natural Language Learning*. ACL, 2009, pp. 147–155.
11. D.M. Blei, "Probabilistic Topic Models," *Comm. ACM*, vol. 55, no. 4, 2012, pp. 77–84.
12. M. Miller and C. Gupta, "Mining Traffic Incidents to Forecast Impact," *Proc. ACM SIGKDD Int'l Workshop on Urban Computing*, 2012, pp. 33–40; http://doi.acm.org/10.1145/2346496.2346502.
13. F.C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Text Analysis in Incident Duration Prediction," *Transportation Research Part C: Emerging Technologies*, vol. 37, Dec. 2013, pp. 177–192.
14. F.C. Pereira et al., "Why So Many People? Explaining Non-Habitual Transport Overcrowding with Internet Data," submitted for publication.
15. C. Collins, S. Hasan, and S. Ukkusuri, "A Novel Transit Riders' Satisfaction Metric: Riders' Sentiments Measured from Online Social Media Data," *J. Public Transportation*, vol. 16, no. 2, 2013; http://shar.es/FVoOb.
16. B. Pender et al., "Social Media Utilisation during Unplanned Passenger Rail Disruption What's Not to Like?" *Proc. Australasian Transport Research Forum 2013*, 2013; www.atrf.info/papers/2013/index.aspx.
17. A.K. McCallum, "Mallet: A Machine Learning for Language Toolkit," 2002; http://mallet.cs.umass.edu.
18. Alias-i, "Lingpipe 4.1.0," 2008; http://alias-i.com/lingpipe.

Submission draft version. This article was published in IEEE Intelligent Systems, vol. 29, no. 1, pp. 76-80.

19. R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks*, Univ. of Malta, 2010, pp. 45–50; http://is.muni. cz/publication/884893/en.
20. M. Ben-Akiva et al., “Traffic Simulation with DynaMIT,” *Fundamentals of Traffic Simulation*, Springer, 2010, pp. 363–398.
21. Mahmassani, Hani S. "Dynamic network traffic assignment and simulation methodology for advanced system management applications." Networks and Spatial Economics 1.3-4 (2001): 267-292.
22. T. Hou et al., “Calibration of Traffic Flow Models under Adverse Weather and Application in Mesoscopic Network Simulation Procedures,” *Proc. Transportation Research Board 92nd Ann. Meeting*, 2013, no. 13-5359.
23. M. Ben-Akiva and S. Lerman, *Discrete Choice Analysis: Theory and Application to Predict Travel Demand*, vol. 9, MIT Press, 1985.
24. J.A. Lopes, “Traffic Prediction for Unplanned Events on Highways,” PhD dissertation, Instituto Superior Tecnico (IST), 2012.
25. D. Nam and F. Mannering, “An Exploratory Hazard-Based Analysis of Highway Incident Duration,” *Transportation Research Part A: Policy and Practice*, vol. 34, no. 2, 2000, pp. 85–102.
26. C. Antoniou, H.N. Koutsopoulos, and G. Yannis, “Dynamic Data-Driven Local Traffic State Estimation and Prediction,” *Transportation Research Part C: Emerging Technologies*, vol. 34, 2013, pp. 89–107; www.sciencedirect.com/science/article/pii/S0968090X13001137.
27. Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017*, tech. report, 2012.
28. International Data Corp., *Worldwide Quarterly Mobile Phone Forecast*, tech. report, 2013.

Francisco C. Pereira *is a senior research scientist in the Singapore-MIT Alliance for Research and Technology, Singapore and a professor in the Departamento de Engenharia Informática, University of Coimbra, Portugal. Contact him at camara@smart.mit.edu.*

Ana L.C. Bazzan *is an associate professor in the Instituto de Informática at the Universidade Federal do Rio Grande do Sul. Contact her at bazzan@inf.ufrgs.br.*

Moshe Ben-Akiva *is the Edmund K. Turner Professor of Civil and Environmental Engineering at the Massachusetts Institute of Technology. Contact him at mba@mit.edu.*

Submission draft version. This article was published in IEEE Intelligent Systems, vol. 29, no. 1, pp. 76-80.