Published in final edited form as:

Proc Hum Factors Ergon Soc Annu Meet. 2014 September ; 58(1): 2107–2111. doi:

10.1177/1541931214581443

Variations on a theme: Topic modeling of naturalistic driving data

Eleese McLaurin¹, Anthony D. McDonald¹, John D. Lee¹, Nazan Aksan², Jeffrey Dawson², Jon Tippin², and Matthew Rizzo²

¹University of Wisconsin-Madison

²University of Iowa

Abstract

This paper introduces Probabilistic Topic Modeling (PTM) as a promising approach to naturalistic driving data analyses. Naturalistic driving data present an unprecedented opportunity to understand driver behavior. Novel strategies are needed to achieve a more complete picture of these datasets than is provided by the local event-based analytic strategy that currently dominates the field. PTM is a text analysis method for uncovering word-based themes across documents. In this application, documents were represented by drives and words were created from speed and acceleration data using Symbolic Aggregate approximation (SAX). A twenty-topic Latent Dirichlet Allocation (LDA) topic model was developed using words from 10,705 documents (real-world drives) by 26 drivers. The resulting LDA model clustered the drives into meaningful topics. Topic membership probabilities were successfully used as features in subsequent analyses to differentiate between healthy drivers and those suffering from Obstructive Sleep Apnea.

Introduction

Naturalistic driving data captured by in-vehicle data recorders (IVDRs) may uncover subtle causative factors in crashes, which are often impossible to identify through standard crash investigations and reconstructions. These data also facilitate comparisons between driving in hazardous conditions and normal driving conditions. Naturalistic driving studies using IVDRs have targeted a range of populations of interest including teenagers, older adults, and truck drivers (Jovanis, Aguero-Valverde, Wu, & Shankar, 2011) and have resulted in an increasing number of datasets that remain to be fully explored. Naturalistic driving studies in uncontrolled real-world settings present a challenge relative to traditional experimentation because the amount of data produced by naturalistic driving studies overwhelms traditional data analysis methods. This challenge motivates an exploration of novel data reduction and analysis methods.

The large amount of data commonly collected during naturalistic driving studies makes comprehensive analysis prohibitive without some type of data reduction. Often this leads researchers to extract and analyze a subset of the data, such as safety-related events (Jovanis et al., 2011; Klauer, Dingus, & Neale, 2006; Shankar, Jovanis, Aguero-Valverde, & Gross, 2008). This approach is helpful, but it excludes the broader context of the drive that may reveal causal factors in the event. This context can be obtained by using data reduction

methods that allow analysis of the entire dataset. One promising data reduction method is Symbolic Aggregate approximation (SAX) time-series analysis (Lin, Keogh, Wei, & Lonardi, 2007; McDonald et al., 2013). SAX compresses the data by averaging windows of the data while retaining the original structure of the data. The input to SAX is a continuous source of data, such as speed or acceleration, and the output is a series of letters.

The conversion of continuous data to symbols facilitates the use of natural language processing algorithms to analyze driving data. These algorithms have been applied with considerable success in other domains and might uncover insights regarding driving behavior that might otherwise be obscured by the volume of naturalistic driving data. One such algorithm is probabilistic topic modeling (PTM). PTM is an unsupervised machine learning method designed to use the words in a body of documents to determine the topics that comprise each of the documents. Each topic is a set of words specified by a probability distribution over the vocabulary of the documents (Blei, 2012). The use of PTM has extended beyond text analyses to subjects such as population genetics, computer vision and human activity recognition (Blei, 2012; Farrahi & Gatica-Perez, 2011; Huynh, Fritz, & Schiele, 2008). This study introduces the use of PTM for the analysis of naturalistic driving data, to differentiate between healthy drivers and drivers with Obstructive Sleep Apnea (OSA). The following sections describe the naturalistic data, SAX time-series analysis, and PTM. The application of PTM to naturalistic driving data demonstrates how it might complement traditional data reduction and data analysis procedures. The paper concludes with a discussion of the utility of this approach.

Naturalistic Driving Data

The data used in this work originate from an ongoing NIH-funded study that is focused on evaluating the effects of adherence to positive airway pressure (PAP) therapy on drivers with Obstructive Sleep Apnea (OSA). Participants include drivers with OSA and healthy drivers who are matched by age, gender, and other demographic factors to a driver from the OSA group. Currently 106 participants have begun the 3.5 month data collection process, 71 OSA drivers (mean age 46), and 35 healthy drivers (mean age 45).

Throughout the study, participants' own vehicles were equipped with vehicle instrumentation that recorded GPS information, speed, and three-axis acceleration at 10Hz. Data from participants with OSA were collected for two weeks prior to receiving therapy and for three months following the start of therapy. The data were partitioned into individual drive files defined by ignition engagement and disengagement. Driving data, PAP usage and sleep behavior were monitored daily. The present analyses focus on speed and acceleration data as these are particularly robust and potentially diagnostic of driver behaviors (Fuller, 2005).

Sax Time-Series Analysis

The driving dataset contains 42,602 drives-over 14,000 hours of driving-with speed and acceleration collected at 10 Hz. This volume of data makes manual inspection intractable, and the variety of driving situations makes simple summaries inappropriate. Symbolic Aggregate approximation (SAX) time-series analysis provides a method for reducing this

data into a more tractable form by converting the data to words that can be analyzed with natural language processing techniques. The SAX method reduces continuous data by replacing small segments of data with the mean of the segment. The means are converted to symbols via bins of the measure, which are often defined by quantiles of the normal distribution (Lin, Keogh, Wei, & Lonardi, 2007). The input to SAX is a continuous source of data and the output is a series of letters.

Figure 1 shows the steps of SAX reduction for a ten second segment of speed data. The top plot shows the raw data at 10 Hz. The middle plot shows the mean reduction step to reduce the data to 1 Hz (1 letter per second), and introduces the labeled bins on the y axis. The bottom plot shows the final symbolic conversion overlaid on the raw data. The letter sequence, “dfghhhhhii,” is the final output of SAX for this segment of speed data. SAX time-series analysis was applied separately to speed, lateral acceleration (lat. acc.), and longitudinal acceleration (lon. acc.). Table 1 shows the bins used for this analysis. The most common acceleration letter was “e”, indicating no rapid change in speed and no rapid change in heading. Common speed letters include “a” (stopped), “e” (city driving), and “h” (highway driving). After applying SAX, words corresponding to one second of driving data were created by combining letters from the speed, lat. acc., and lon. acc., as shown in Figure 2.

Probabilistic Topic Modeling

Probabilistic topic modeling can uncover meaningful, thematic structure in a set of documents (Blei, 2012). The exact thematic structure and meaning of each topic depends on the documents. For example, a topic modeling can be applied to the thousands of abstracts published in the Human Factors and Ergonomics journals to identify topics that define the field, such as automation, displays, and driving (Lee, 2014). Once identified, these themes reveal connections between seemingly disparate elements of the field, as well as trends in topic popularity. Topic modeling offers a “far reading” global overview of many documents that complements the more familiar “close reading” of the local details within a particularly document (Mimno, 2012). In a similar fashion, topic modeling can describe the general context of driving to make the detailed analysis of events more meaningful.

Latent Dirichlet Allocation models

The simplest probabilistic topic modeling algorithm is Latent Dirichlet Allocation (LDA). LDA defines topics as distributions over a fixed vocabulary of words. Each document, or set of words, exhibits multiple topics to various degrees (Blei, Ng, & Jordan, 2003). Thus a fitted LDA model provides a set of likelihoods associated with pairs of words and topics, and pairs of documents and topics. LDA models are fit by learning the parameters of the topic distributions.

LDA models are generative models, meaning they fit a joint probability distribution over data to produce a model. This joint distribution is defined over the number of topics, the topic proportions for each document, the topic assignment for each word in each document, and the observed words. Given a fixed number of topics, documents, and an observed set of

words from the documents, the remaining parameters can be estimated in a way that maximizes the likelihood of the model given the data.

LDA has three input parameters: the number of topics, the vocabulary, and Dirichlet distribution parameters. The number of topics can be determined by iteratively fitting models and assessing the perplexity of the model, which is an assessment of overall model fit across the topics (Grun & Hornik, 2011). The vocabulary size is often reduced using the TFIDF metric which eliminates very common or very rare words from the documents. The goal of this reduction is to both expedite the analysis and target it towards the discovery of meaningful themes. In natural language processing this reduction prevents the discovery of themes related to words that might be rare across articles and also themes related to common words, such as “a” or “the” that have no substantive meaning in a larger context. The distributional parameters associated with the Dirichlet distribution, denoted alpha and beta, control the degree to which terms or documents are associated with a single topic or multiple topics. These parameters can be used to coerce terms and documents to represent a single topic or allow them to represent multiple topics.

PTMs and driving

The application of PTMs to driving requires a shift in perspective from documents and words to drives and driving words. SAX identifies driving words for each drive, which in turn, become a direct analog for a document. Applying a probabilistic topic model to this data will produce a set of drive themes. The meaning of these themes will depend, to some extent, on the number of topics in the model. For example, a model with few topics would identify highly salient drive characteristics, such as the long periods of high speed driving that characterizes highway driving. A model with a larger number of topics might identify more subtle characteristics, such as those that relate to impaired highway driving. Topic modeling can benefit current analysis methods by both revealing patterns of behavior that differentiate conditions of interest and by identifying covariates for traditional statistical analyses.

To assess and define the benefits of PTM for driving we fit a topic model to a sample of 26 drivers, thirteen healthy drivers, and thirteen drivers with OSA, using the R topicmodels package (Grun & Hornik, 2011). Each healthy driver was matched to an OSA driver of similar age, gender, marital status, and level of education. The sample contained 10,705 drives, approximately 410 per driver. This sample was selected to minimize variance due to gender, and age. Twenty topics were found to minimize perplexity for this dataset. The total vocabulary size was 284, no words were removed during preprocessing as all the words were determined to be meaningful. Alpha and beta were estimated automatically as a part of the model fitting. To improve solution stability, ten twenty-topic models were fitted and the model with the lowest log-likelihood was selected.

Results and Discussion

The fitted topic model produces two probability distributions of interest. The first describes the likelihood (or strength) of each topic in each drive and the second describes the strength of each word relative to each topic. Together these probability distributions can be used to

generate a definition of each topic. For the model fitted in this analysis, the topics can be described as clusters of speed ranges and acceleration ranges. A schematic of the topics is shown in Figure 3. As an example of the topic contents, Figure 4 shows the most representative drive for topics 5, 6, and 7, based on the magnitude of the topic probabilities. These three topics form the cluster for speed “d” (20-30 mph) with topic 5 representing drives with little to no change in acceleration (word e_e_d), Topic 6 representing drives with changes in lat. acc. (words e_f_d & e_d_d) and topic 7 representing drives with changes in lon. acc. (words f_e_d & d_e_d).

The results illustrate the ability of the topic model to identify commonly understood and similar, but difficult to define drives. This combination is important because it differentiates topic models from more simple analyses based on thresholding (Dozza, Bargman, & Lee, 2012). While these methods may identify the drives associated with topic 5, it would be difficult for them to distinguish the more complex patterns identified in topics 6 and 7.

In addition to assessing the driving themes revealed by the topic model, we wanted to assess whether the topics can distinguish between drivers with and without OSA. For the first part of this assessment, the mean topic probabilities for each condition were compared. Based on the mean topic probabilities, topic 7 was identified as a particularly distinguishing topic. (OSA: $M0.026$, $SD0.052$; Healthy: $M0.055$, $SD0.074$). Figure 5 shows a boxplot of the distribution of mean values for topic 7 and suggests that topic probabilities can distinguish between drivers with and without OSA.

To systematically identify which topics might distinguish OSA drivers from healthy drivers, the topic probabilities were used as features in a random forest machine learning algorithm. The algorithm was fit to the data and assessed using 10-fold cross validation.

The algorithm succeeded in differentiating OSA and healthy drivers with an Area Under Curve (AUC) of 0.80, sensitivity of 0.78, specificity of 0.66 and mean accuracy of 72.7%. Cohens Kappa was 0.45, indicating moderately good agreement between the algorithm predictions and the clinical diagnosis of OSA.

The random forest machine learning algorithm fit was substantially better than a logistic regression (AUC=0.69), indicating that OSA is identified by a complex combination of topic probabilities and not a linear function of them. The most important variables for the random forest model were topics 7, 6, and 5. These topics represent drives at lower speeds (under 30 mph). The least important variables were topics 16, 20, and 19. These topics represent drives at high speeds (over 60 mph). These differences suggest that OSA-related driving behavior is more distinctive at lower speeds. These results are promising, however further analysis is needed to determine what specific characteristics of these drives are linked to the effects of OSA.

Limitations

This study explored topic modeling for analyzing naturalistic driving data. As demonstrated in this study, topic models provide robust descriptions of the driving data that can serve as

useful inputs to subsequent analyses. As with all unsupervised machine learning methods, topic model results require external validation and interpretation.

Currently, the most accurate validation method for topic models is visual inspection of the SAX terms and drives associated with each topic to assess whether the topics are meaningful or spurious. For models with a large number of topics, this process can be tedious, however the process is comparable to that used to verify a similar number of threshold partitions, and much less onerous than manual inspection of all the drives. Furthermore, once the model is validated, it can be used to describe future datasets.

An important consideration in this data analysis approach is the selection of input parameters for SAX and the topic models. These parameters require an optimization process that could improve results; however, the processing time for such an optimization process can be substantial. The definition of the words is another important consideration. In this study, combining SAX letters together over a one-second window created words. While this method is efficient, more descriptive words might be identified with more sophisticated segmentation methods (Lin et al., 2007; McDonald et al., 2013). Finally, the insight afforded by this modeling process may be enhanced by using a larger and more diverse database of drivers.

Future work

Further exploration will examine alternative word definitions, optimization of parameters, and models that include n-grams, or patterns of multiple words, rather than single word analyses. More generally, a supervised approach to topic modeling might prove particularly useful in distinguishing between types of drivers. This study fit topic models in an unsupervised manner independently of the driver type. Topic models could be fit in a supervised manner such that topics are identified to differentiate between drivers with and without OSA, enabling the resulting topics to differentiate driver types more efficiently.

Conclusion

Naturalistic driving data present an unprecedented opportunity to understand driving behavior. To fully realize the potential of such data, researchers need to employ novel data analysis methods that provide a more complete picture than the event-based approach that dominates much of the naturalistic data analyses. This paper presents one such method: Probabilistic Topic Modeling. This method offers value by identifying meaningful patterns in large volumes of data, while dramatically reducing the size of the dataset. In this paper, we showed a drive can be meaningfully described according to the probability of 20 topics—a reduction of three orders of magnitude for a 12-minute drive. These topics differentiate between driver conditions and enhance the general understanding of driving data.

References

- Blei DM. Probabilistic topic models. *Communications of the ACM*. 2012; 55(4):77.doi: 10.1145/2133806.2133826
- Blei DM, Ng A, Jordan M. Latent dirichlet allocation. *The Journal of Machine Learning Research*. 2003; 3:993–1022. Retrieved from <http://dl.acm.org/citation.cfm?id=944937>.

- Dozza M, Bargman J, Lee JD. Chunking: A procedure to improve naturalistic data analysis. *Accident Analysis & Prevention*. 2012; doi: 10.1016/j.aap.2012.03.020
- Farrahi K, Gatica-Perez D. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology*. 2011; 2(1):1–27. DOI: 10.1145/1889681.1889684
- Fuller R. Towards a general theory of driver behaviour. *Accident Analysis & Prevention*. 2005; 37(3): 461–72. DOI: 10.1016/j.aap.2004.11.003 [PubMed: 15784200]
- Grun B, Hornik K. topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*. 2011; 40(13):1–30.
- Huynh, T., Fritz, M., Schiele, B. Discovery of activity patterns using topic models; Proceedings of the 10th International Conference on Ubiquitous Computing - UbiComp '08; 2008. p. 10
- Jovanis PP, Aguero-Valverde J, Wu K, Shankar V. Analysis of Naturalistic Driving Event Data Omitted-Variable Bias and Multilevel Modeling Approaches. *Transportation Research Record: Journal of the Transportation Research Board*. 2011; 2236:49–57.
- Klauer, S., Dingus, T., Neale, V. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. 2006. Retrieved from <http://trid.trb.org/view.aspx?id=786825>
- Lee JD. Text analysis for topics and trends in Human Factors and Ergonomics publications. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (in Review). 2014
- Lin J, Keogh E, Wei L, Lonardi S. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*. 2007; 15(2):107–144. DOI: 10.1007/s10618-007-0064-z
- McDonald AD, Lee JD, Aksan NS, Dawson JD, Tippin J, Rizzo M. The Language of Driving. *Transportation Research Record: Journal of the Transportation Research Board*. 2013; 2392:22–30. DOI: 10.3141/2392-03
- Mimno D. Computational historiography. *Journal on Computing and Cultural Heritage*. 2012; 5(1):1–19. DOI: 10.1145/2160165.2160168
- Shankar V, Jovanis PP, Aguero-Valverde J, Gross F. Analysis of Naturalistic Driving Data: Prospective View on Methodological Paradigms. *Transportation Research Record: Journal of the Transportation Research Board*. 2008; 2061:1–8. -1. DOI: 10.3141/2061-01

Figure 1.

SAX time-series reduction for a segment of speed data. The three graphs show the raw data, the reduction step, and the symbolic conversion step respectively.

Figure 2. The conversion between SAX output and PTM input

Figure 3. Schematic of topic clusters

Figure 4. SAX speed, lat. acc. and lon. acc. for the most representative drive for Topics 5, 6, and 7

Figure 5. Mean comparison for Topic 7

Table 1**Bins for SAX**

Lat. & Lon. Acc.		Speed	
Letter	Range	Letter	Range
a	< -0.6g	a	0 mph
b	(-0.6) to (-0.4)g	b	<10 mph
c	(-0.4) to (-0.3)g	c	10-20 mph
d	(-0.3) to (0.05)g	d	20-30 mph
e	(-0.05) to (0.05)g	e	30-40 mph
f	0.05 to 0.3g	f	40-50 mph
g	0.3 to 0.4g	g	50-60 mph
h	0.4 to 0.6g	h	60-70 mph
i	> 0.6g	i	>70 mph