Text Mining and Topic Modeling of Compendiums of Papers from Transportation Research Board Annual Meetings

Subasish Das, Xiaoduan Sun, and Anandi Dutta

The collective knowledge system has been advancing rapidly in the recent past. The digitalization of information in many online media—such as blogs, social media, articles, webpages, images, audios, and videos—provides an unprecedented opportunity for the extraction and identification of a knowledge trend. Prominent journal and conference proceedings usually contain extensive amounts of textual data that can be used to examine the research trends for various topics of interest and to understand how this research has helped in the advancement of a subject such as transportation engineering. The exploration of the unstructured contents in journal or conference papers requires sophisticated algorithms for knowledge extraction. This paper presents text mining techniques to analyze compendiums of papers published from TRB annual meetings, the largest and most comprehensive transportation conferences in the world. Topic models are algorithms designed to discover hidden thematic structure from massive collections of unstructured documents. This study used a popular topic model, latent Dirichlet allocation, to reveal research trends and interesting histories of the development of research by analyzing 15,357 compendiums of papers from 7 years (2008 to 2014) of TRB annual meetings.

The rise of the Internet and digital gadgets has evolved publication media into an abundant source of usability of information as part of a collective knowledge system. This escalation provides unprecedented opportunities to identify and investigate new knowledge and its applications. The prospect and correct use of new age information are evolving issues in theory and practice. Moreover, the amount of extractable information has been growing exponentially. This information explosion requires sophisticated tools and models to make the dissemination of new knowledge effectual.

Text mining has gained popularity among researchers, as it helps to automate knowledge extraction from unstructured large textual data. The rapid advancement in machine learning and natural language processing has introduced probabilistic-framework text-mining models called topic models. These models are based on the idea that documents are mixtures of topics in which a topic is a probability distribution over words. Among topic models, latent Dirichlet allocation (LDA) models are widely used. An LDA topic model is a Bayesian mixture model for discrete data in which topics are uncorrelated.

TRB organizes the largest and most comprehensive annual transportation conference in the world. Established in 1920 as the National Advisory Board on Highway Research, TRB provides a mechanism for the exchange of information and research results about every aspect of transportation, with a focus on highway transportation research and development. The mission of TRB is to promote innovation and progress in transportation through research. In an objective and interdisciplinary setting, TRB facilitates the sharing of information on transportation practice and policy by researchers and practitioners, stimulates research and offers research management services that promote technical excellence, provides expert advice on transportation policy and programs, and disseminates research results broadly and encourages their intuitive implementation.

The research papers published in the TRB annual meeting compendiums are peer-reviewed by hundreds of TRB committees, and a small percentage of compendium papers are accepted for publication in the *Transportation Research Record.* The large textual data in TRB compendium papers require sophisticated tools for knowledge dissemination. However, text mining on TRB compendium papers had never been performed. This study aims to focus on a topic discovery system to reveal the implicit knowledge in TRB compendium papers.

LITERATURE REVIEW

Text mining, the process of deriving high-quality information from text, is having a wider range of applications in many fields. With the increasing power of computers and programming software, text mining can now explore any large amount of textual data within a limited time and limited resource allocation for easy-to-understand knowledge. As a newer branch in scientific data analysis, text mining is growing quickly. Semantic analysis of the textual data was widely used to facilitate many applications, such as user interest modeling (*1*), sentiment analysis (*2*), content exploration (*3–5*), event tracking (*6*), citizen–government relationships (*7–9*), news retrieving (*10*), prediction of stock market variations (*11*), management of natural disasters (*12*), understanding of epidemical diseases (*13*), and characterization of electoral processes (*14*).

Topic modeling is a type of statistical model for discovering the unstructured topics that occur in a collection of documents. Blei wrote

S. Das, Texas A&M Transportation Institute, Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135. X. Sun, Civil Engineering Department, and A. Dutta, Center for Advanced Computer Studies, University of Louisiana at Lafayette, Lafayette, LA 70504. Corresponding author: S. Das, s-das@tti.tamu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2552, Transportation Research Board, Washington, D.C., 2016, pp. 48–56.
DOI: 10.3141/2552-07

a general introductory article on topic modeling with an emphasis on LDA (*15*). Research trend analysis using topic models has been conducted in several studies. In 2008, Hall et al. performed a study to investigate the development of ideas in the scientific field by using LDA (*16*). Paul and Girju used LDA to develop a novel classifier to classify research papers on the bases of topic and language (*17*). Recently, Cui et al. used topic models to explore trends in cancer research (*18*). Research work by Berry and Castellanos explains the recent focus on text mining methods (*19, 20*). The research team of the current study compiled detailed bibliographies (with abstracts of the papers) on text mining and topic modeling in two web pages (*21, 22*).

METHODOLOGY

TRB's activities regularly engage more than 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia. State transportation departments and federal agencies, including the component administrations of the U.S. Department of Transportation (DOT), support this program. TRB organizes 5,000 presentations in nearly 750 sessions and workshops covering a board area of transportation. The objective of the current research is to investigate how data mining can be helpful in extracting and identifying new knowledge and applications of that knowledge from TRB annual meeting publications. Table 1 shows the number of papers published annually in the TRB compendiums of papers and an increase in the number of papers accepted for the compendiums over time; for example, the number of papers in a compendium increased 46% from 2009 to 2014. To accomplish the goal of the current research, the research team used two data mining methods: text mining and topic modeling.

Text Mining

Text mining is an applied method that originated from a more generic scientific branch called data mining or knowledge discovery (KD), which is the scientific process of identifying valid, original, important, and ultimately interpretable patterns in both structured and unstructured data. Knowledge discovery in text (KDT) or text mining can be viewed as a multifaceted process that comprises all activities from document collection to interpretable knowledge extraction. KDT uses methods like data mining, information retrieval, supervised and unsupervised machine learning, and computational semantics. Information retrieval from databases (in this case, TRB papers) through pattern recognition helps identify contributing factors or trends in the associated tasks. Text mining mainly deals with collections of unstructured textual data rather than structured databases.

With text mining methods, users assume that keywords represent compact information in documents. Keyword extraction uses a method of natural language processing to identify particular word–term tags that are combined by various machine-learning algorithms. Another text mining area of particular interest in many studies is the co-occurrence of particular phrases and terms. For example, high frequency of the term "congestion" would indicate the nature of the document's particular interest. The high occurrence of "congestion" with the term "minimal" would indicate a rather different nature of the document's interest. In text mining, a "corpus" represents a collection of text documents. A corpus is an abstract concept, and it can have several implementations in parallel. After developing a corpus, users can clean the textual contents by removing redundant words, phrases, numbers, and punctuation marks to make the content less noisy.

The main information in the compendiums of papers collected from TRB includes the following attributes:

- Publication year,
- Title of the paper,
- Abstract,
- Author names,
- First-author affiliation,
- Review committee code, and
- Review committee name.

Table 2 lists the top 10 review committees and clearly reveals both the contemporary issues of concern and the focus areas in transportation. The current research also investigated the demographics of the research community. Figure 1 shows the top 10 first-author affiliations in the published TRB compendia of papers. Even though approximately 90% of TRB annual meeting attendees are from the United States, two of the top 10 first-author affiliations are from outside the country.

The text corpus, the collection of texts, was created on the basis of the annual compendium of papers. Each of the seven years of the compendium was stored in a consecutively numbered document,

TABLE 1 Number of TRB Papers by Year

Year	Number of Compendium Papers
2008	697
2009	1,993
2010	2,143
2011	2,312
2012	2,539
2013	2,758
2014	2,915

TABLE 2 Top 10 Review Committees

No.	Reviewing Committee's Name	TRB Code	Papers in Compendium
1	Traveler Behavior and Values	ADB10	512
2	Traffic Flow Theory and Characteristics	AHB45	403
3	Safety Data, Analysis, and Evaluation	ANB20	331
4	Transportation Demand Forecasting	ADB40	309
5	Traffic Signal Systems	AHB25	280
6	Pedestrians	ANF10	247
7	Transportation and Air Quality	ADC20	243
8	Transportation in the Developing Countries	ABE90	243
9	Transportation Network Modeling	ADB30	242
10	Bicycle Transportation	ANF20	232

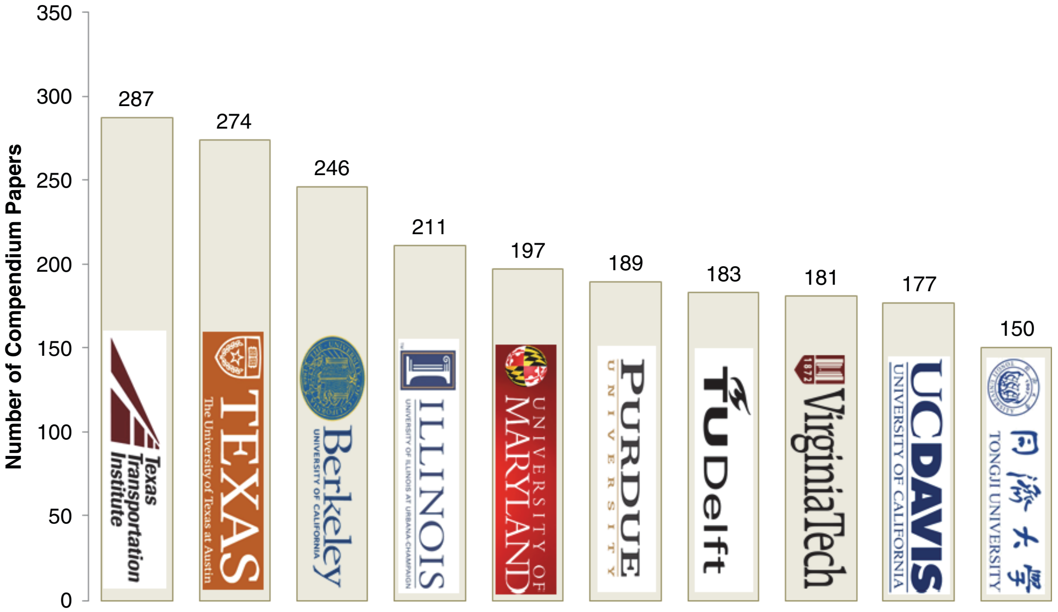

FIGURE 1 Top 10 first-author affiliations.

beginning with TRB 2008. In conducting text mining, the research team used two data groups: paper abstract and paper title. For abstracts, the team used a random sample of 3,000 representative abstracts to mitigate computational delay. For paper titles, the team used all 15,357 papers for analysis.

Table 3 lists the basic outputs generated from these two groups (paper abstracts and paper titles). "Sparsity" is a standard measure that represents the rare occurrence of terms in the whole document. In the initial analysis, sparsity was around 62% for both groups. This study removed sparse terms (i.e., terms occurring only in very few documents) to narrow the matrix of terms dramatically without losing significant relationships inherent to the matrix. The number of terms was reduced to 1,084 and 2,293 for the groups of paper titles and paper abstracts, respectively, after the removal of sparse terms.

One of the most common tasks in text mining is determining the most frequently cited terms in a corpus. Figure 2 illustrates the most frequently cited terms found in the combined corpus from all paper titles. The five most frequently cited terms are "model"–"models"–"modeling," "traffic," "analysis," "pavement," and "evaluation"–"evaluations"–"evaluating". Figure 3 shows the heat map of the 20 most frequently cited terms in paper titles. The darker color indicates a higher percentage of usage, while the lighter color indicates a lower percentage. One is not surprised to see that some terms (such as "models"–"modeling," "evaluation"–"evaluating," and "traffic") remain popular over the 7 years and that usage frequency of some terms (such as "urban," "pavement," "concrete," and "safety") changes over time. Particularly, pavement-related research shows a clear decline in recent years.

TABLE 3 Analysis Before and After Removal of Sparse Terms

Variable	Paper Titles	Sample of Abstracts
Analyzed Papers	15,357	3,000
All Terms		
Nonsparse/sparse entries	30,371/52,334	52,337/86,228
Terms	11,815	19,795
Sparsity	63%	62%
Maximal term length	34	45
After Removing Sparse Terms (14%)		
Nonsparse/sparse entries	7,588/0	16,051/0
Terms	1,084	2,293
Sparsity	0%	0%
Maximal term length	22	21

Figure 4 illustrates the most frequently cited terms found in the combined corpus from the sample group of paper abstracts. The five most frequently cited terms are "model"–"models"–"modeling," "data"–"data set"–"database," "traffic," "travel"–"travels"–"travelers," and "vehicle"–"vehicles"–"vehicular." Figure 5 shows the heat map of the yearly weights of the 20 most frequently cited terms in the abstracts of the randomly sampled papers. Usage of some terms (such as "transit," "information," "asphalt," and "crash") changes over time.

Most-cited terms exhibit slight differences between the groups of paper titles and abstracts. For example, in the group of paper titles, "data" and its related terms ("data set," "database," etc.) is ranked second in the group of abstracts, pushing "traffic" to the third position. "Model" or "modeling" is the most frequently used word in both groups. This difference is easily interpreted, as most paper abstracts contain a brief introduction of the used data.

A word cloud is another way of visualizing the most frequent terms in unstructured documents. In performing a general word cloud, this study developed comparison word clouds to visualize the research trends over time. If $p_{a,b}$ is the rate at which word a occurs in document b, and p_b is the average rate across documents ($\Sigma_b p_{a,b}/n$), where n is the number of documents. In comparison clouds, the size of each word is mapped to its maximum deviation ($\max_a(p_{a,b} - p_b)$), and its angular position is determined by the document in which

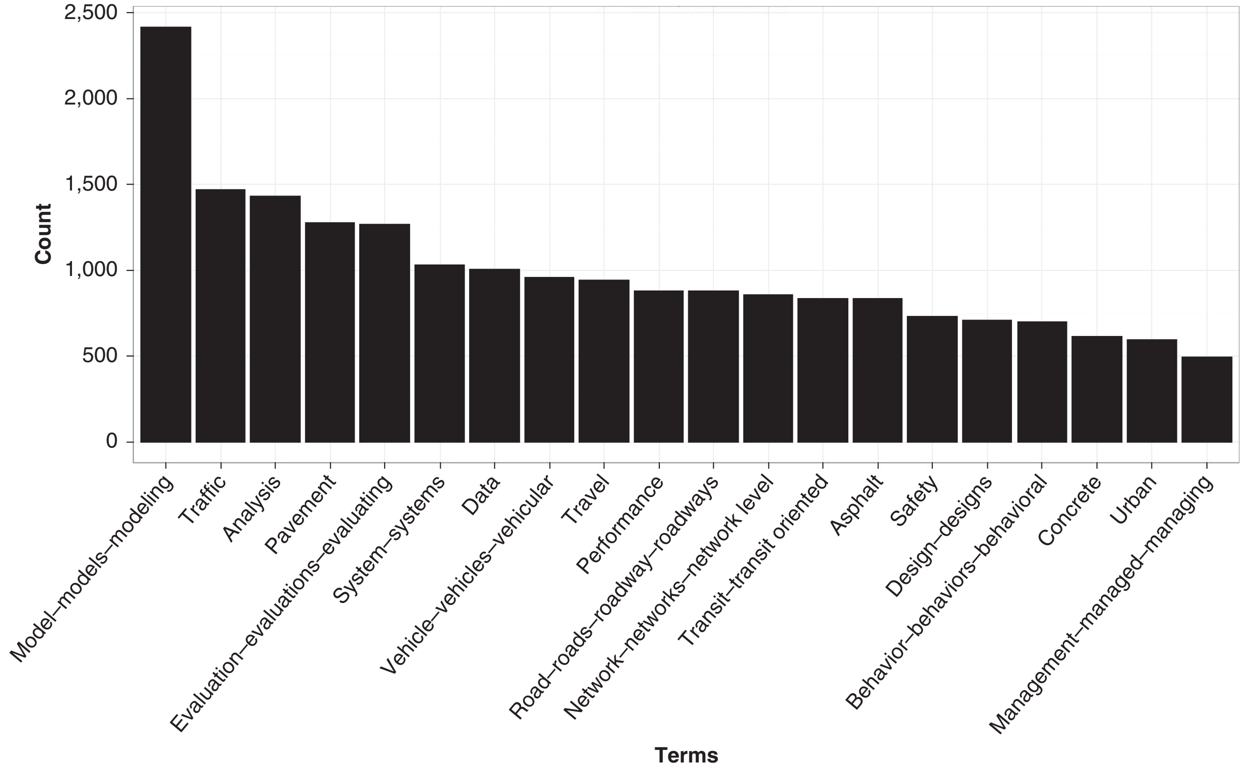

FIGURE 2 Most frequently used terms in paper titles.

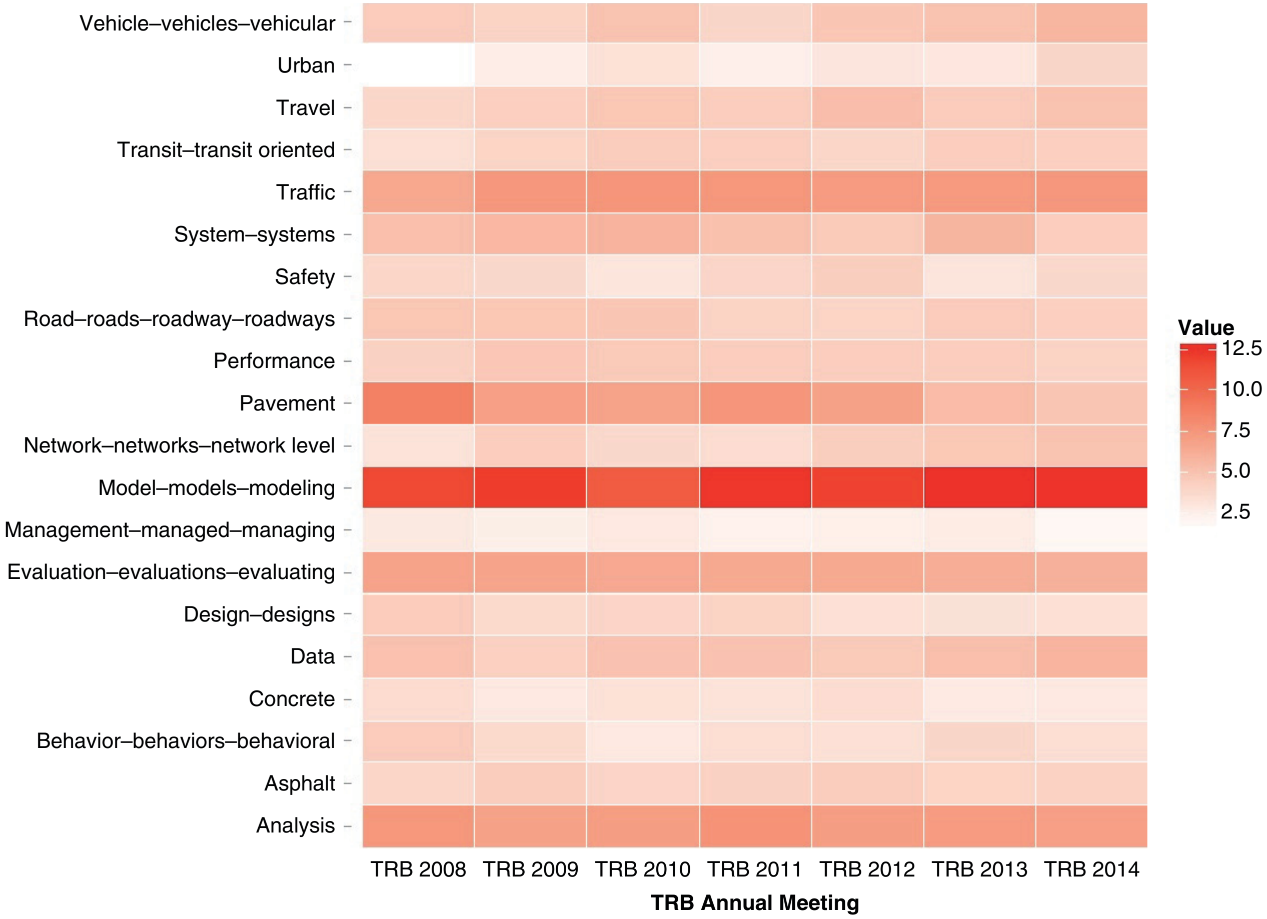

FIGURE 3 Heat map of most frequently used terms in paper titles.

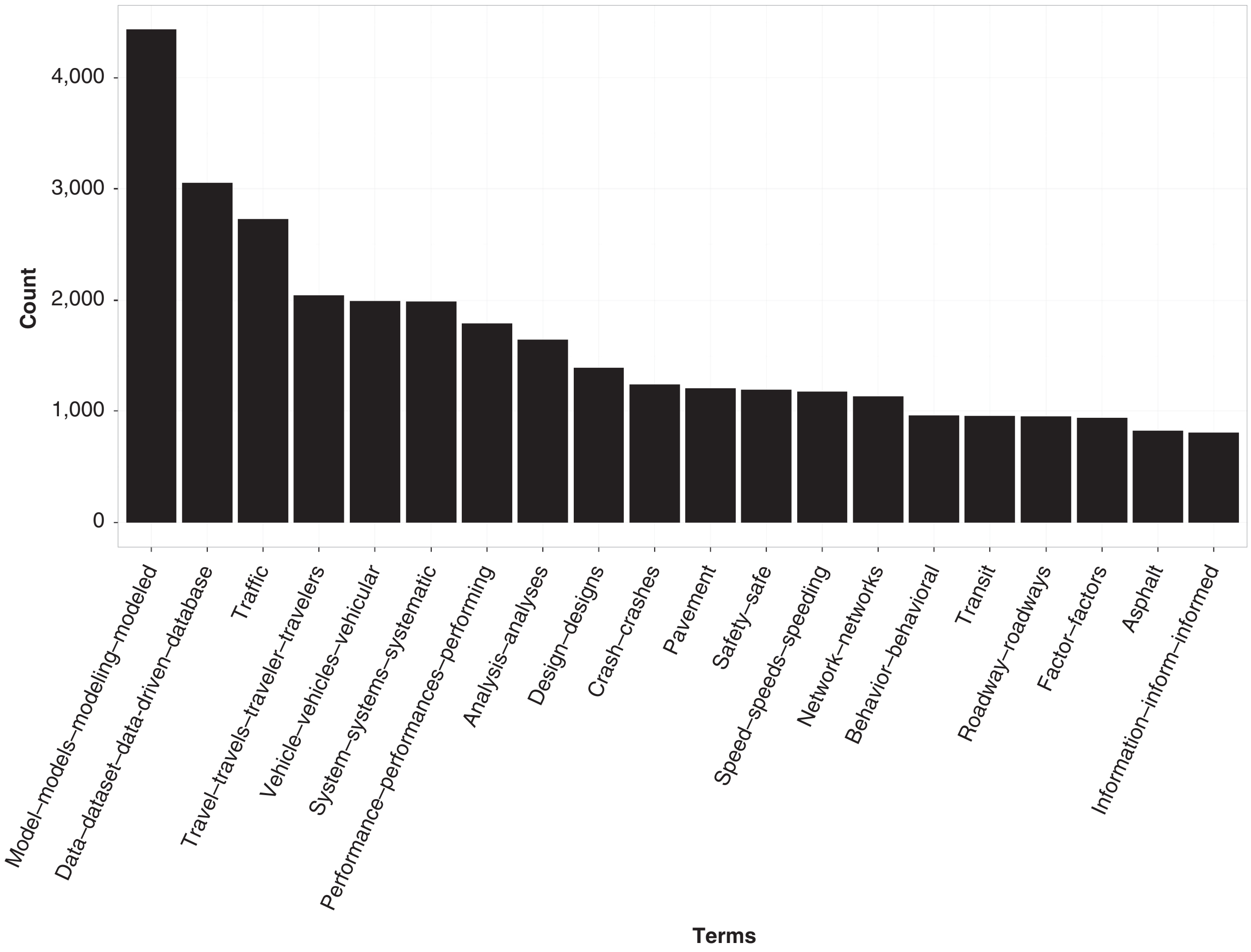

FIGURE 4 Most frequently used terms in paper abstracts.

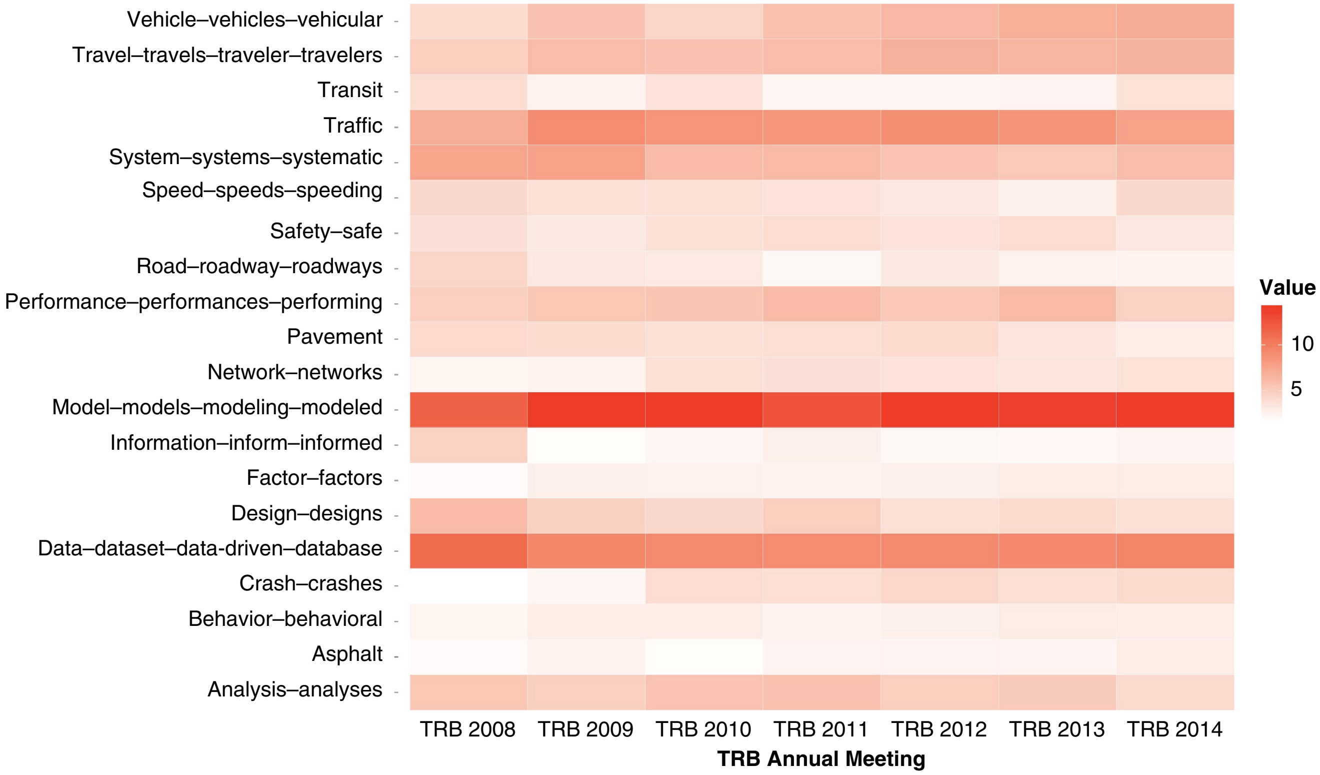

FIGURE 5 Heat map of most frequently used terms in paper abstracts.

that maximum occurs. For this analysis, the research team excluded TRB 2008 papers for tidy visualization. The team used combined groups of TRB papers from 2009 and 2010, from 2011 and 2012, and from 2013 and 2014. Figure 6 illustrates comparison clouds for TRB paper titles from 2009 to 2014, starting with 2009 and 2010 papers, for which the size of the word indicates the usage (bigger words imply more frequent usage). Figure 7 shows comparison clouds for paper review committees.

The popular word or words may vary year by year. For example, the most significant terms are "planning," "pavement," and "modeling" for the groups of TRB papers from 2009 and 2010, 2011 and 2012, and 2013 and 2014, respectively. The newer research trend in social media is visible in the term "social" in the group of 2013 and 2014 TRB papers. In Figure 7, the most significant terms are "bituminous," "management," and "characteristics" for the groups of TRB papers from 2009 and 2010, 2011 and 2012, and 2013 and 2014, respectively.

Table 4 shows the correlation between the words of the four most frequently cited terms in both paper titles and sampled abstracts. To make this research replicable, codes used in this study are made available on a web page (*23*).

To see that "model" is completely associated with "development" or that "traffic" always appears with "evaluating" in paper titles comes as no surprise. In sampled abstracts, "model" is highly correlated with "capacity," "effects," and "results."

LDA Modeling

LDA is an autonomous way of discovering topics in unstructured documents. Several authors have used LDA in the development of topic models. A convenient source for a short introduction on the theoretical development of LDA is a study by Blei et al. (*24*).

LDA modeling assumes that documents are represented as random mixtures over latent topics in which each topic is characterized by a distribution over words. One can consider a document as a sequence of N words denoted by $t = (t_1, t_2, \dots, t_N)$, where t_n is the nth word in the sequence. A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \dots, V\}$. Words represent unit-basis vectors that have a singular component equal to one and the rest of the components equal to zero. Thus, by using superscripts to denote components, the vth word in the vocabulary is represented by a V-vector w such that $t^v = 1$ and $t^u = 0$ for $u \neq v$. A corpus is a collection of M documents denoted by $T = \{t_1, t_2, \dots, t_M\}$.

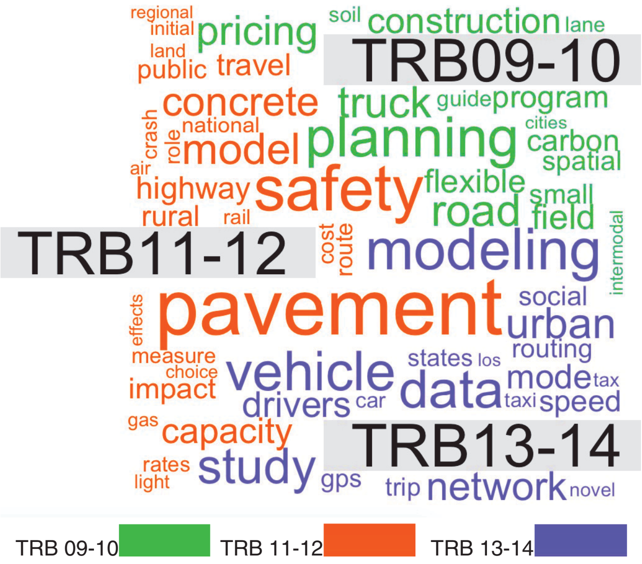

FIGURE 6 Comparison word clouds of paper title terms.

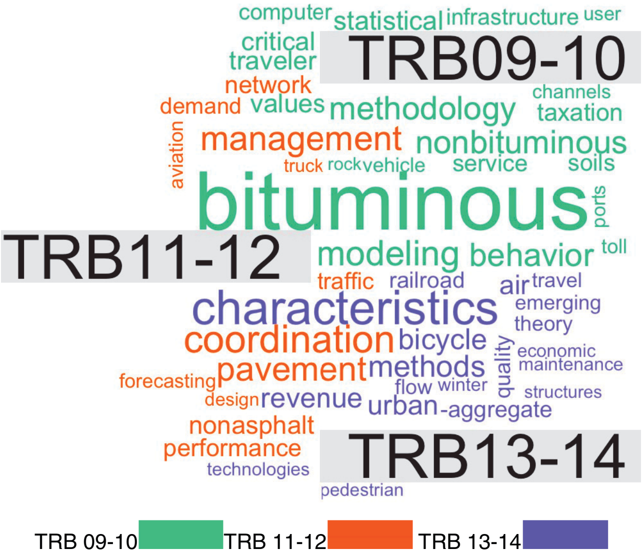

FIGURE 7 Comparison word clouds of review committees.

This study's research team used LDA models with the following steps for each document w in a corpus T:

1. Consider the number of the multinomial $N \sim \text{Poisson}(\varphi)$.
2. Consider the parameter of class distribution $\theta \sim \text{Dirichlet}(\alpha)$. Here, α is the parameter of a Dirichlet distribution (DD) over the hidden classes.

TABLE 4 Correlations Between Terms

Paper Titles		Sampled Abstracts	
Model		Model	
Development	1	Capacity	0.99
Vehicles	1	Effects	0.99
Dynamic	0.98	Results	0.99
Traffic		Traffic	
Evaluating	1	Differences	1
Analysis	0.99	Quantified	1
Performance	0.99	Identified	1
		Improved	0.99
Analysis		Analysis	
Evaluating	1	Bridge	0.99
Traffic	0.99	Ensure	0.99
		Networks	0.98
Pavement		Pavement	
Preservation	0.98	Increases	0.99
Deformation	0.98	Possible	0.99
Overlays	0.97		
Sensors	0.96		

3. For each of N words t_n
   – Consider a topic z_n ~ multinomial(θ).
   – Choose a word t_n from $p(t_n|z_n)$, a multinomial probability conditioned on the topic z_n.

This study considered several assumptions. First, the dimensionality k of the DD is assumed known and fixed, and then the word probabilities are parameterized by a $k \times V$ matrix β, where $\beta_{ij} = p(t_j = 1|z_i = 1)$ that is treated as a fixed quantity needing to be estimated. N is independent of the other data-generating variables (θ and z). Thus, it is an ancillary variable, and randomness is not further considered.

A k-dimensional Dirichlet random variable θ for a specific α can be written as follows:

$$P(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1} \qquad (1)$$

Here, the parameter α is a k-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the gamma function. Given the parameters α and β, the joint distribution of a latent class mixture θ, a set of N latent classes z, and a set of N features t is given by Equation 2:

$$P(\theta, z, t|\alpha,\beta) = P(\theta|\alpha)\prod_{n=1}^{N}P(z_n|\theta)P(t_n|z_n,\beta) \qquad (2)$$

Here, $P(z_n|\theta)$ is θ for each unique i such that z_n^i equals 1. Then integrating over θ and summing over z results in Equation 3:

$$P(t|\alpha,\beta) = \int P(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}P(z_n|\theta)P(t_n|z_n,\beta)\right)d\theta \qquad (3)$$

After taking the product of the marginal probabilities of single documents, one can finally get the probability of a corpus:

$$P(t|\alpha,\beta) = \prod_{t=1}^{M}\int P(\theta_t|\alpha)\left(\prod_{n=1}^{N_t}\sum_{z_{nt}}P(z_{nt}|\theta_t)P(t_{tn}|z_{tn},\beta)\right)d\theta_t \qquad (4)$$

The thematic nature of any document collection evolves over time, and therefore explicitly modeling the dynamics of the hidden correlation is important. For developing the topic models for two groups (titles and abstracts) of text documents, this study used open source R software package topic models (*25*). Topic extraction from a text corpus is fundamental to many topic analysis tasks. In this analysis, latent topic models are extended to find the underlying structure of time series in an unsupervised manner (Figures 8 and 9). LDA using bag-of-patterns representation automatically discovered the clusters of topics that are in the unstructured form in the document groups. The generated topics from both document groups are illustrated in Figures 8 through 11. Figure 8 also illustrates the trend of the topics over time.

Figure 10 shows six panels of topics from the group of paper titles, with a set of four tightly co-occurring terms. Topic 1T includes terms like “traffic,” “systems,” “models,” and “data” and clearly indicates traffic data modeling. Topic 2T implies performance modeling. Topic 3T indicates research related to pavement analysis. Topic 4T indicates travel data modeling research. Topics 5T and 6T indicate research on traffic design and traffic application analysis, respectively.

Figure 11 lists eight panels of topics from the group of paper abstracts with a set of six tightly co-occurring terms. Topic 1A includes behavioral divers. Topics 2A and 3A cover research on traffic crash data analysis and traffic network modeling research, respectively. Topic 4A indicates research on project management. Topic 5A indicates research related to nonmotorized mobility options. Topic 6A implies pavement performance analysis. Topics 7A and 8A indicate research on environmental impact and

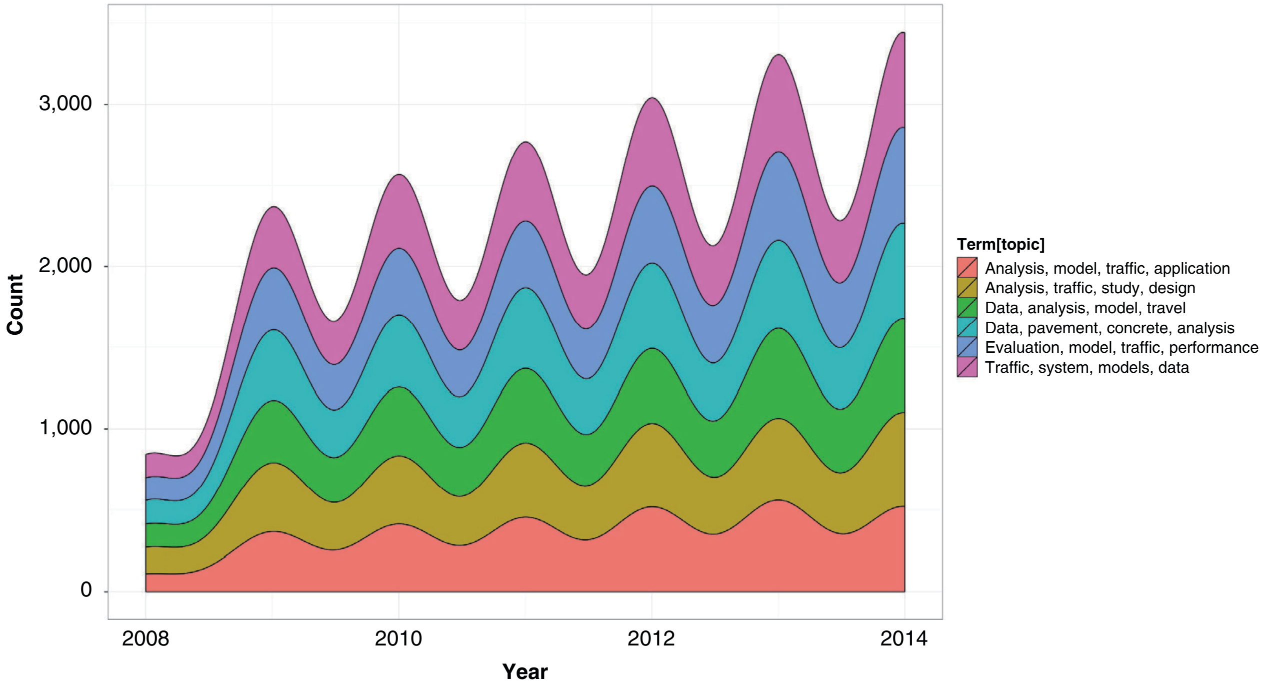

FIGURE 8 Trend of top six topics for paper titles over the years.

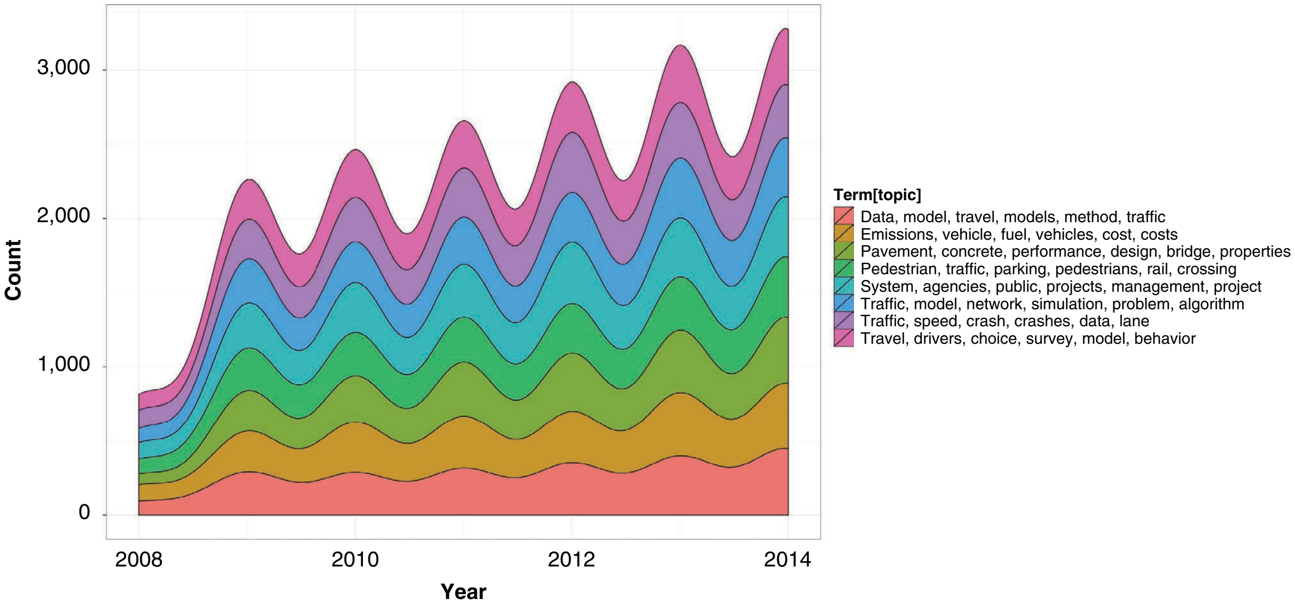

FIGURE 9 Trend of the top eight topics for paper abstracts over the years.

TOPIC 1	TOPIC 2	TOPIC 3
traffic	evaluation	data
systems	model	pavement
models	traffic	concrete
data	performance	analysis

TOPIC 4	TOPIC 5	TOPIC 6
data	analysis	analysis
analysis	traffic	model
model	study	traffic
travel	design	application

FIGURE 10 Top six topics from paper titles.

TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4
travel	traffic	traffic	system
drivers	speed	model	agencies
choice	crash	network	public
survey	crashes	simulation	projects
model	data	problem	management
behavior	lane	algorithm	project

TOPIC 5	TOPIC 6	TOPIC 7	TOPIC 8
pedestrian	pavement	emissions	data
traffic	concrete	vehicle	model
parking	performance	fuel	travel
pedestrians	design	vehicles	models
rail	bridge	cost	method
crossing	properties	costs	traffic

FIGURE 11 Top eight topics from paper abstracts.

modal analysis, respectively. Figure 9 shows the trend of the top eight topics over time.

CONCLUSIONS

By exploring text mining applications in research papers published from TRB annual meetings, the world's most comprehensive transportation conference, this research took advantage of rapidly advancing data analysis techniques to examine research trends and possible emerging knowledge in the area of transportation. The preliminary results have clearly shown that transportation research is truly dynamic. The top topics from the paper abstracts show the recent prominence in behavioral research and safety prediction models. As society moves forward, the critical issues change over time, and solutions to the problems posed by these changing issues require a broad perspective and innovative thinking. The research trends found in the current study will help the TRB community to explore transportation-related research ideas over time.

The current research requires expansion in scope and content. A larger historical data set containing all TRB papers in digital format would permit examination of the research trends and unseen connections more critically. TRB publications are a robust laboratory for multidisciplined research. Another potential source for analysis would be the published papers in the *Transportation Research Record.* Because of extensive scrutiny by peer reviewers, these papers have greater authenticity than those from the compendiums. More advanced topic models like infinite dynamic topic models are potential tools for future research.

ACKNOWLEDGMENTS

The authors deeply appreciate the interest of Mark R. Norman, TRB Director of Development and Strategic Initiatives and former Technical Activities Division director, and Michael DeCarmine, TRB Business Systems Analyst, who provided spreadsheet data on TRB Annual Meeting compendiums of papers, to support the idea for this paper.

REFERENCES

1. Pennacchiotti, M., and S. Gurumurthy. Investigating Topic Models for Social Media User Recommendation. *Proc., 20th International Conference Companion on World Wide Web* (WWW '11), New York, 2011, pp. 101–102.
2. Lin, C., and Y. He. Joint Sentiment–Topic Model for Sentiment Analysis. *Proc., 18th ACM Conference on Information and Knowledge Management* (CIKM '09), Association for Computing Machinery, New York, 2009, pp. 101–102.
3. Duan, J., and J. Zeng. Web Objectionable Text Content Detection Using Topic Modeling Technique. *Expert Systems with Applications,* Vol. 40, 2013, pp. 6094–6104.
4. Martinez-Romo, J., and L. Araujo. Detecting Malicious Tweets in Trending Topics Using a Statistical Analysis of Language. *Expert Systems with Applications,* Vol. 40, 2013, pp. 2992–3000.
5. Waters, R., and J. Jamal. Tweet, Tweet, Tweet: A Content Analysis of Nonprofit Organizations' Twitter Updates. *Public Relations Review,* Vol. 37, 2011, pp. 321–324.
6. Lee, C. Unsupervised and Supervised Learning to Evaluate Event Relatedness Based on Content Mining From Social-Media Streams. *Expert Systems with Applications,* Vol. 39, 2012, pp. 13338–13356.
7. Panagiotopoulos, P., A. Bigdeli, and S. Sams. Citizen–Government Collaboration on Social Media: The Case of Twitter in the 2011 Riots in England. *Government Information Quarterly,* Vol. 31, No. 3, 2014, pp. 349–358.
8. Sobaci, M., and N. Karkin. The Use of Twitter by Mayors in Turkey: Tweets for Better Public Services? *Government Information Quarterly,* Vol. 30, 2013, pp. 417–425.
9. Chatfield, A., H. Scholl, and U. Brajawidagda. Tsunami Early Warnings via Twitter in Government: Net-Savvy Citizens' Co-Production of Time-Critical Public Information Services. *Government Information Quarterly,* Vol. 30, 2013, pp. 377–386.
10. Hong, S. Online News on Twitter: Newspapers' Social Media Adoption and Their Online Readership. *Information Economics and Policy,* Vol. 24, 2012, pp. 69–74.
11. Bollen, J., H. Mao, and X. Zeng. Twitter Mood Predicts the Stock Market. *Journal of Computer Science,* Vol. 2, No. 1, 2011, pp. 1–8.
12. Sakaki, T., M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. *Proc., 19th International Conference on the World Wide Web* (WWW '10). Association for Computing Machinery, New York, 2010, pp. 851–860.
13. Culotta, A. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. *Proc., 1st Workshop on Social Media Analytics* (SOMA '10). Association for Computing Machinery, New York, 2010, pp. 115–122.
14. Borondo, J., A. Morales, J. Losada, and R. Benito. Characterizing and Modeling an Electoral Campaign in the Context of Twitter: 2011 Spanish Presidential Election as a Case Study. *Chaos,* Vol. 22, No. 2, 2012.
15. Blei, D. Probabilistic Topic Models. *Communications of the ACM.* Vol. 55, No. 4, 2012.
16. Hall, D., D. Jurafsky, and C. Manning. Studying the History of Ideas Using Topic Models. *Proc., 2008 Conference on Empirical Methods in Natural Language Processing,* Honolulu, Hawaii, Oct. 2008.
17. Paul, M., and R. Girju. Topic Modeling of Research Fields: An Interdisciplinary Perspective. *Proc., International Conference RANLP (Recent Advances in Natural Language Processing),* Borovets, Bulgaria, 2009, pp. 337–342.
18. Cui, M., Y. Liang, Y. Li, and R. Guan. Exploring Trends of Cancer Research Based on Topic Model. Presented at 1st International Workshop on Semantic Technologies, Changchun, China, March 9–12, 2015.
19. Berry, M., and M. Castellanos. *Survey of Text Mining: Clustering, Classification, and Retrieval.* Springer, New York, 2004.
20. Berry, M., and M. Castellanos. *Survey of Text Mining II.* Springer, New York. 2008.
21. Bibliography of social media research. http://subasish.github.io/pages/TRB2016/textm.html. Accessed July 17, 2015.
22. Bibliography of topic mining research. http://subasish.github.io/pages/TRB2016/topicm.html. Accessed July 17, 2015.
23. Web page for topic modeling paper. http://subasish.github.io/pages/TRB2016/topic_pap.html. Accessed July 17, 2015.
24. Blei, D., A. Ng, and M. J. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research,* Vol. 3, 2003, pp. 993–1022.
25. Grun, B., and K. Hornik. topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software,* Vol. 40, No. 13, 2011, pp. 1–30.

The Standing Committee on Library and Information Science for Transportation peer-reviewed this paper.