Text Mining and Topic Modeling on Compendium Papers from Transportation Research Board Annual Meetings

Subasish Das, Ph.D.¹
Associate Transportation Researcher
Texas A&M Transportation Institute
Texas A&M University System
3135 TAMU
College Station, TX 77843-3135
Email: s-das@tti.tamu.edu
Phone: 979-845-9958
Fax: 979-845-6006

Xiaoduan Sun, Ph.D. & PE
Professor
Civil Engineering Department
University of Louisiana
Lafayette, LA 70504
Email: xsun@louisiana.edu
Phone: 337-739-6732
Fax: 337-739-6688

Anandi Dutta
Ph.D. Candidate
Center for Advanced Computer Studies
University of Louisiana at Lafayette
Lafayette, LA 70504
Email: axd1329@louisiana.edu

Word Count: 7,000 including 11 figures and 4 tables

¹Corresponding Author

Submitted to the 95th TRB Annual Meetings for Presentation and Publication under Committee Library and Information Science for Transportation (ABG40)
ABSTRACT

The collective knowledge system is advancing rapidly in the age of the Internet. The digitalization of information in many online medias- like blogs, social media, articles, webpages, images, audios, and videos- provides an unprecedented opportunity for us to extract and identify new knowledge. Prominent journal and conference proceedings usually contain extensive amounts of textual data that can be used to examine the research trends on various interested topics and to understand how these researches have helped the advancement of a particular subject such as transportation engineering. The exploration of the unstructured contents in journal or conference papers requires sophisticated knowledge extraction algorithms. Topic models are algorithms designed to discover the main theme or trend from massive collections of unstructured documents. This paper utilizes text mining techniques to analyze compendium papers published by Transportation Research Board (TRB) Annual Meetings, the largest and most comprehensive transportation conference in the world. A total of 15,357 compendium papers from seven years (2008-2014) of TRB Annual Meeting conference proceedings were analyzed by using a popular topic modeling, named latent Dirichlet allocation (LDA), to reveal research trends and interesting research development histories.

Key words: Topic modeling, latent Dirichlet allocation, text mining.
INTRODUCTION

The rise of the Internet and digital gadgets has evolved publication media into an abundant source of information in digital format as part of a collective knowledge system, which provides unprecedented opportunities for us to extract, identify and investigate new knowledge and its applications. The prospect and correct use of digitalized information is an evolving issue in theory and practice. The amount of information in digital form has been growing exponentially. Sophisticated tools and models are needed for researchers to analyze this vast amount of information so that the new knowledge can be disseminated effectively.

The rapid development in machine learning and natural language processing (NLP) has offered a probabilistic framework for term frequency occurrence models, named as topic models, in documents based upon the idea that documents are mixtures of topics where a topic is a probability distribution over words. Among different topic models, latent Dirichlet allocation (LDA) models are widely used. An LDA topic model is a Bayesian mixture model for discrete data where topics are uncorrelated.

The Transportation Research Board (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 implementation.

The research papers published by the TRB annual meeting compendiums are peer reviewed by hundreds of TRB committees and a small percentage of compendium papers are accepted for publication by the TRB journal. It is interesting to investigate research trends and applications by analyzing the TRB compendium papers with currently evolving data mining tools such as text mining and topic modeling. In this paper, we focused on a topic discovery system to reveal the implicit knowledge present 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 massive amount of textual data for easy-to-understand knowledge. As a newer branch in scientific data analysis, text mining is growing quickly. There are two good literature review papers on text mining that explore both theoretical aspect and application-oriented methodologies [1-2]. Semantic analysis of the textual data was widely used to facilitate many applications such as user interest modeling [3], sentiment analysis [4], content exploration [5-7], event tracking [8], citizen-government relations [9-11], news retrieving [12], prediction of stock market variations [13], the management of natural disasters [14], the understanding of epidemical diseases [15] and the characterization of electoral processes [16].
Topic modeling is a type of statistical model for discovering the unstructured topics that occur in a collection of documents. Blei wrote a general introductory article on topic modeling with an emphasis on latent Dirichlet allocation (LDA) [17]. Research trend analysis by 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 [18]. Paul and Girju used LDA to develop a novel classifier to classify research papers based on topic and language [19]. Recently, Cui et al. used topic model to explore trends of cancer research [20]. The researchers of the current study compiled detailed bibliographies (with abstracts of the papers) on text mining and topic modeling in two webpages [21-22]. Interested readers can surf through these two web pages for more detailed reviews on text mining and topic modeling.

METHODOLOGY

TRB’s various 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, support this program. More than 5,000 presentations in nearly 750 sessions and workshops were made covering a broad area of transportation. The objective of this research is to investigate how data mining can be helpful in retraction and identifying new knowledge and applications of that knowledge from TRB Annual Meeting publications. As a first step, the data from 15,357 compendium papers from 2008-2014 TRB Annual Meetings were collected as shown in Table 1. It is seen that the number of accepted compendium papers increased over the years; for example, a 46% increase in number of compendium papers is seen from 2009 to 2014. Accomplishing the research goal, two data mining methods were used in this study: text mining and topic modeling.

TABLE 1. Number of TRB Papers by Year

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Compendium Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>697</td>
</tr>
<tr>
<td>2009</td>
<td>1,993</td>
</tr>
<tr>
<td>2010</td>
<td>2,143</td>
</tr>
<tr>
<td>2011</td>
<td>2,312</td>
</tr>
<tr>
<td>2012</td>
<td>2,539</td>
</tr>
<tr>
<td>2013</td>
<td>2,758</td>
</tr>
<tr>
<td>2014</td>
<td>2,915</td>
</tr>
</tbody>
</table>

Text Mining

Text mining is an applied method that originated from a more generic scientific branch called data mining or knowledge discovery. Knowledge discovery (KD) is the scientific process of identifying valid, original, important, and ultimately interpretable patterns in unstructured data. Knowledge discovery in text (KDT) or text mining can be viewed as a multi-stage 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. Extraction of useful information from databases (in this
case, TRB papers) through pattern recognition helps identify contributing factors or trends in associated tasks. Text mining mainly deals with the collections of unstructured textual data rather than from structured databases.

In text mining methods, it is assumed that keywords represent compact information in documents. Keyword extraction uses a natural language processing method to identify particular word/term tags that is combined with various machine learning algorithms. Another particular interest is seen in co-occurrence of particular phrases and terms. This co-occurrence would be a point of interest in many studies. For example, high frequency of the term *congestion* would indicate the nature of the document’s particular interest. If the occurrence of *congestion* with another term *minimal* is high, it 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 they can be several implementations in parallel. After developing a corpus, users can clean the textual contents by removing redundant words, numbers, punctuations, etc.

The main information on TRB compendium papers collected from TRB authority contains the following attributes:
- Publication year
- Title
- Abstract
- First author affiliation
- Review committee’s code
- Review committee’s name

Table 2 lists the top ten review committees, which clearly reveals the contemporary concerned issues and highlighting areas in transportation. The research also investigated the demographics of the research community. Figure 1 shows the top ten first author affiliations in TRB compendium paper publications. It is interesting to note that even though approximately 90% of TRB annual meeting attendees are from the U.S., two of the top ten first author affiliations are outside of the country.

### Table 2. Top Ten Review Committees

<table>
<thead>
<tr>
<th>No.</th>
<th>Reviewing Committee's Name</th>
<th>TRB Code</th>
<th>Papers Appeared in Compendium</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Traveler Behavior and Values</td>
<td>ADB10</td>
<td>512</td>
</tr>
<tr>
<td>2</td>
<td>Traffic Flow Theory and Characteristics</td>
<td>AHB45</td>
<td>403</td>
</tr>
<tr>
<td>3</td>
<td>Safety Data, Analysis and Evaluation</td>
<td>ANB20</td>
<td>331</td>
</tr>
<tr>
<td>4</td>
<td>Transportation Demand Forecasting</td>
<td>ADB40</td>
<td>309</td>
</tr>
<tr>
<td>5</td>
<td>Traffic Signal Systems</td>
<td>AHB25</td>
<td>280</td>
</tr>
<tr>
<td>6</td>
<td>Pedestrians</td>
<td>ANF10</td>
<td>247</td>
</tr>
<tr>
<td>7</td>
<td>Transportation and Air Quality</td>
<td>ADC20</td>
<td>243</td>
</tr>
<tr>
<td>8</td>
<td>Transportation in the Developing Countries</td>
<td>ABE90</td>
<td>243</td>
</tr>
<tr>
<td>9</td>
<td>Transportation Network Modeling</td>
<td>AFK50</td>
<td>242</td>
</tr>
<tr>
<td>10</td>
<td>Bicycle Transportation</td>
<td>ANF20</td>
<td>232</td>
</tr>
</tbody>
</table>
The text corpus, collection of texts, was created based on the annual compendium papers. Seven years of TRB compendium papers were stored in seven documents named TRB2008, TRB2009, TRB2010, TRB2011, TRB2012, TRB2013, and TRB2014 respectively. In conducting text mining, two specific data groups were considered: Paper Title and Paper Abstract. For paper abstracts, 3,000 representative abstracts were randomly selected for analysis to mitigate computational delay. For the group of paper titles, all 15,357 paper tiles were used in the analysis.

Table 3 lists the basic outputs generated from these two groups (paper titles and paper abstracts). For both of the groups, the sparsity is around 62%. The study removed sparse terms, i.e., the terms occurring only in very few documents, to narrow down the matrix of the terms dramatically without losing significant relations inherent to the matrix. The number of terms reduced to 1,084 and 2,293 for the groups of paper titles and paper abstracts, respectively, after the removal of sparse terms.

**TABLE 3. Analysis of Before and After Sparse Term Removal**

<table>
<thead>
<tr>
<th></th>
<th>Paper Titles</th>
<th>Sample of Abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyzed Papers</td>
<td>15,357</td>
<td>3,000</td>
</tr>
<tr>
<td><strong>All Terms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-sparse/Sparse entries</td>
<td>30,371/52,334</td>
<td>52,337/86,228</td>
</tr>
<tr>
<td>Terms</td>
<td>11,815</td>
<td>19,795</td>
</tr>
<tr>
<td>Sparsity</td>
<td>63%</td>
<td>62%</td>
</tr>
<tr>
<td>Maximal Term length</td>
<td>34</td>
<td>45</td>
</tr>
<tr>
<td><strong>After Removing Sparse Terms (14%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-sparse/Sparse entries</td>
<td>7,588/0</td>
<td>16,051/0</td>
</tr>
<tr>
<td>Terms</td>
<td>1,084</td>
<td>2,293</td>
</tr>
<tr>
<td>Sparsity</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Maximal Term length</td>
<td>22</td>
<td>21</td>
</tr>
</tbody>
</table>
One of the most common tasks in text mining is to find out the most frequently cited terms in the corpus. Figure 2 illustrates the most frequently cited terms found in the combined corpus from all paper titles. It is interesting to note that the top five most frequently cited terms are: model/models/modeling, traffic, analysis, pavement, and evaluations/evaluations/evaluating. Figure 3 shows the heatmap of the top 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. It is not surprising to see some terms remain popular over the seven years such as models/modeling, evaluation/evaluating and traffic, and usages of some terms do change over time such as urban, pavement, concrete, and safety. Particularly, pavement related research shows a clear decline in the recent years.

FIGURE 2. Most frequently used terms in paper titles.

FIGURE 3. Heatmap of most frequently used terms in paper titles.
Figure 4 illustrates the most frequently cited terms found in the combined corpus from the sampled paper abstracts. The top five most frequently cited terms are: model/models/modeling, data/dataset/database, traffic, travel/travels/travelers, and vehicle/vehicles/vehicular. Figure 5 shows the heatmap of the yearly weightage of the top 20 most frequently cited terms in the randomly sampled paper abstracts. Usages of some terms do change over time such as transit, information, asphalt, and crash.

**FIGURE 4.** Most frequently used terms in paper abstracts.

**FIGURE 5.** Heatmap of most frequently used terms in paper abstracts.
There are slight differences in most cited terms between the groups of paper titles and abstracts. For example, comparing with the group of paper titles, *data* and anything related to data (*dataset*, *database* and etc.) is ranked number two in the group of abstracts, pushing the word *traffic* into number three. *Model* or *modeling* is still the most frequently used word in both groups. This is easily interpreted as most paper abstracts contain a brief introduction of the used data.

Wordcloud is another way of visualizing the most frequent terms in unstructured documents. In performing a general wordcloud, this study developed comparison wordclouds to visualize the research trends over the years. If $p_{a,b}$ is the rate at which word $a$ occurs in document $b$, and $p_b$ is the average across documents ($\sum_b p_{a,b}/n$), 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 where that maximum occurs. For this particular analysis, we excluded TRB 2008 papers for better visualization. Three groups were formed: TRB papers in 2009-2010, TRB papers in 2011-2012, and TRB papers 2013-2014. Figure 6 illustrates comparison clouds for TRB paper titles from 2009 to 2014 starting with TRB09-10 papers and working forward, where size of the word indicates the usage (bigger size words infer more frequent usage). Figure 7 shows comparison clouds for paper review committees.

It seems that the popular word or words may vary year by year. For example, the most significant terms are *planning*, *pavement*, and *modeling* for TRB 2009-2010, TRB 2011-2012, and TRB 2013-2014, respectively. The newer research trend in social media is visible by the term *social* in TRB 2013-2014 papers. In Figure 6, the most significant terms are *bituminous*, *management*, and *characteristics* for TRB 2009-2010, TRB 2011-2012, and TRB 2013-2014, respectively.
Table 4 shows the correlation between the words of the top four most frequently cited terms in both paper title and sampled abstracts. A more detailed dataset is provided in a webpage dedicated for interested readers [23].

**TABLE 4. Correlations Between the Terms**

<table>
<thead>
<tr>
<th>Model</th>
<th>Paper Titles</th>
<th>Sampled Abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td>1</td>
<td>Capacity</td>
</tr>
<tr>
<td>Vehicles</td>
<td>1</td>
<td>Effects</td>
</tr>
<tr>
<td>Dynamic</td>
<td>0.98</td>
<td>Results</td>
</tr>
<tr>
<td><strong>Traffic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluating</td>
<td>1</td>
<td>Differences</td>
</tr>
<tr>
<td>Analysis</td>
<td>0.99</td>
<td>Quantified</td>
</tr>
<tr>
<td>Performance</td>
<td>0.99</td>
<td>Identified</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Improved</td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluating</td>
<td>1</td>
<td>Bridge</td>
</tr>
<tr>
<td>Traffic</td>
<td>0.99</td>
<td>Ensure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Networks</td>
</tr>
<tr>
<td><strong>Pavement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preservation</td>
<td>0.98</td>
<td>Increases</td>
</tr>
<tr>
<td>Deformation</td>
<td>0.98</td>
<td>Possible</td>
</tr>
<tr>
<td>Overlays</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>Sensors</td>
<td>0.96</td>
<td></td>
</tr>
</tbody>
</table>
It is not surprising to see the word *model* is completely associated with *development*, or that the word *traffic* always appears with *evaluating* in paper titles. In sampled abstracts, the word *model* is highly correlated with *capacity*, *effects* and *results*.

**Topic Modeling**

*Latent Dirichlet allocation (LDA)*

Latent Dirichlet allocation (LDA) is an autonomous way of discovering topics in unstructured documents. Several authors have used LDA in topic model development. The study by Blei et al. is excellent in describing the theoretical development of LDA [24].

It is assumed that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. We can consider a document as a sequence of $N$ words denoted by $t = (t_1, t_2, ..., t_n)$, where $t_n$ is the $n$th word in the sequence. A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, ..., V\}$. We represent words using unit-basis vectors that have a single component equal to all other components equal to zero. Thus, using superscripts to denote components, the $v$-th 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 $T = \{t_1, t_2, ..., t_M\}$.

For this study we developed an LDA model application with the following steps for each document $w$ in a corpus $T$:

1. Consider the number of the multinomial, $N \sim \text{Poisson}(\phi)$
2. Consider the parameter of class distribution, $\theta \sim \text{Dirichlet}(\alpha)$. Here, $\alpha$ is the parameter of Dirichlet Distribution (DD) over the hidden classes.
3. For each of $N$ words $t_n$:
   - Consider a topic $z_n \sim \text{Multinomial}(\theta)$.
   - Choose a word $t_n$ from $p(t_n | z_n)$, a multinomial probability conditioned on the topic $z_n$.
   
Several assumptions were made in the study: First the dimensionality $k$ of the DD is assumed as known and fixed, and then the word probabilities are parameterized by a $k \times V$ matrix $\beta$ 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 all the other data generating variables ($\theta$ and $z$). Thus, it is an ancillary variable and randomness is not further taken care of.

A $k$-dimensional Dirichlet random variable $\theta$ for a specific $\alpha$ can be written as:

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

Here, the parameter $\alpha$ is a $k$-vector with components $\alpha_i > 0$, and where $\Gamma(\cdot)$ is the Gamma function. Given the parameters $\alpha$ and $\beta$, the joint distribution of a latent class mixture $\theta$, a set of $N$ latent classes $z$, and a set of $N$ features $t$ is given by:

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

Here, $p(z_n | \theta)$ is $\theta$ for each unique $i$ such that $z_n^i$ equals to 1. After, integrating over $\theta$ and summing over $z$:
After taking the product of the marginal probabilities of single documents, we finally get the probability of a corpus:

\[
P(T | \alpha, \beta) = \prod_{t=1}^{M} P(\theta_t | \alpha) \left( \prod_{n=1}^{N_t} P(z_n | \theta_t) P(t_n | z_n, \beta) \right) \, d\theta_t
\]

For developing the topic models for two different groups (titles and abstracts) of text documents, we used open source R software package \textit{topicmodels} [25]. Here all of the TRB papers were considered as a single document per group. Topic extraction from text corpus is the fundamental of many topic analysis tasks. In this analysis, latent topic models are extended to find the underlying structure of time series in an unsupervised manner (\textit{Figure 9} and \textit{Figure 11}). 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 of the document groups are illustrated in \textit{Figure 8-11}.

As shown in \textit{Figure 8}, topic 1 includes terms like traffic, systems, models, and data. It clearly indicates traffic data modeling. The second topic infers performance modeling. The third topic indicates research related to pavement analysis. Topic 4 indicates travel data modeling research. Topic 5 and Topic 6 indicate research on traffic design and traffic application analysis, respectively. The trend of the topics over the years is shown in \textit{Figure 9}.
FIGURE 9. Trend of top six topics for paper titles over the years.

As shown in Figure 10, topic 1 includes behavioral research on drivers. The second topic and third topic covers research on traffic crash data analysis, and traffic network modeling research, respectively. The fourth topic indicates research on project management. The fifth topic indicates research related to non-motorized mobility options. The sixth topic infers pavement performance analysis. Topic 7 and topic 8 indicate research on environmental impact, and modal analysis, respectively. The trend of these eight topics over the years is shown in Figure 11.

FIGURE 10. Top eight topics from paper abstracts.
FIGURE 11. Trend of top eight topics for paper abstracts over the years.

CONCLUSION

By exploring text mining applications in research papers published by TRB Annual Meetings, the most comprehensive transportation conference in the world, this research took advantage of rapidly advancing data analysis techniques to examine the research trend and possible emerging knowledge in the area of transportation. The preliminary results have clearly shown that transportation research is truly a dynamic area. The top topics from the paper abstracts show the recent prominence in behavioral research and safety prediction models. As human society moves forward, the critical issues change over time from emphasizing mobility only to a more inclusive perspective for problem solutions. The research trend from the current study will help the TRB community to identify the exploration of transportation related research ideas over time.

The current research should be expanded in scope and content. A larger historical dataset containing all TRB papers in digital format should be analyzed to examine the research trends and unseen connections as text mining and topic modeling are becoming ever more efficient with increased computing power. TRB publications are a robust laboratory for such research. Many potential improvements have been identified by the research team; such as developing better interactive visualization on the research findings. The researchers are currently developing an interactive web tool on the findings of this research.

This exploratory research demonstrates the feasibility of using the text mining and topic modeling techniques for synthesis study with massive historical textual data. After performing the initial analysis, a more detailed framework for further studies is in progress.
ACKNOWLEDGEMENTS

TRB Technical Activities Division Director Mr. Mark R. Norman, and TRB Program Officer Michael DeCarmine provided the data in spreadsheet form containing information of TRB Annual Meeting Compendium papers to us. Their interest in our idea and support is deeply appreciated.

REFERENCES


