Utilizing Topic Modeling Techniques to Identify the Emergence and Growth of Research Topics in Engineering Education

Aditya Johri, G. Alan Wang, Xiaomo Liu and Krishna Madhavan
ajohri@vt.edu, alanwang@vt.edu, xiaomliu@vt.edu, cm@purdue.edu

Abstract – **In this paper we present findings from a project that used topic modeling and associated techniques to chart the emergence and growth of research topics in engineering education research over 9 years, from 2000-2008. As a field engineering education research has undergone significant changes over the past decade. There has been an increase in the number of scholars and practitioners involved in the field, particularly those that are applying rigorous research principles to advance understanding of engineering education. In such a circumstance, it is important to understand the topics, approaches, and ideas that have caught the imagination of people in the community. Since this nature of work has not been done in relation to engineering education research, a significant part of the effort described here is innovative and exploratory in nature where different techniques were tested with the goal to collect a diversity of topics that are of interest to the community. We identify major categories of topics and primary topics of interest to the community. We also identify a lack of engagement with theoretical and analytical ideas as an area of concern.**

Index Terms – Engineering education research, Research topics, Topic modeling.

INTRODUCTION

The field of engineering education has experienced significant maturation as a research enterprise over the past decade. Although the roots of engineering education go back over a century, when *Journal of Engineering Education* published its first issue, in recent years there has been increased focus to improve the empirical foundations of the field and numerous initiatives to develop the field have been created and implemented. Any maturing research field can reap significant advantages from a holistic understanding of its past and current efforts, particularly what topics found favor with researchers earlier, how they have changed, and what are some novel and recurring problems that need to be addressed. Yet, empirical efforts to do so at a smaller scale, such as through interviews and surveys, suffer from problems of bias, validity, and reliability. Recognizing the limitations of other approaches, one of the starting points for this research project was the question: How can we identify and study the exploration of a research field over time, noting periods of gradual development, major ruptures, and most importantly the major topics that have been of interest to member of the field?

Faced with this question, we decided to leverage emerging advances in the data mining and analytics techniques. In particular, our investigation of observing such insights is operated on the unsupervised topic modeling method, Latent Dirichlet Allocation (LDA) [1]. As a comparison, we also extract the most meaningful noun phrase and keyword from documents for topic detection and topic trend analysis. These approaches have been applied to various scientific corpora such as Proceedings of the National Academy of Sciences (PNAS), CiteSeer (a computer and information science paper collection), Proceedings of Neural Information Processing Systems (NIPS), and others, and have shown great capabilities of capturing the dynamics of research. To analyze topics in engineering education we developed a corpus of more than 2,500 articles from two journals and one conference on engineering education: *Journal of Engineering Education* (JEE), *International Journal of Engineering Education* (IJEE), and *Proceedings of Frontiers in Education* (FIE). These publications cover most major research topics across engineering education. We are in the process of adding the *Proceedings of Annual Conference of ASEE* to the corpus as well but our preliminary analysis suggests that the topics remain the same with or without that data.

To perform the analysis we built a system that approaches trends from different perspectives – topics, noun phrases and keywords, and the system provides great flexibility in terms of selection which data to analyze, including its context and time range. The data controller enables the selection of input corpus and the data can be a combination of any journals or any conferences or both. The context controller enables to choose the context for topic analysis. It can either be the title, the abstract or keywords in a paper. The model controller enables to choose the models of extracting topics or concepts in the corpus. It can be either of topic modeling using LDA, noun phrase extraction or keyword extraction. The time controller enables to choose time range to calculate topic trends. It can be either individual years or individual months. The findings indicate that some topics have remained constant over the years but some topics, such as global issues and assessment, have seen significant interest in the past five years. In addition to topic

978-1-61284-469-5/11/$26.00 ©2011 IEEE

modeling, our system and analysis also provides dynamic data on number of papers published over time and other publication characteristics.

RELATED WORK

There have been many studies of the dynamics of scientific research. Using LDA models to capture the trends of topics becomes popular in recent years. Griffiths and Steyvers analyze the hot and cold topics of PNAS articles between 1991 and 2001 as meanings of gaining insights into the dynamics of science [2]. They present a basic analysis based on the post-hoc examination of the estimated probability of a topic to a document produced by the LDA model. Hall et al. apply a similar method in the major conferences of Computational Linguistics from 1978 to 2006 to understand its historical trends [4]. They also introduce a model of the diversity of ideas, topic entropy, which is able to show the topic diversities of difference conferences. Wang and McCallum extend the original LDA model by directly incorporate the topic changes over time [8]. Unlike some prior work mentioned previously, their model parameterizes a continuous distribution over time associated with each topic. Their experiments on several real-world data sets show the discovery of more salient topics that are clearly localized in time than the plain LDA model. Despite of the popularity of using LDA family models for trend analysis, other methods based on noun phrases and keywords are proposed and proved to be effective. For example, Jo et al. address the problem of detecting topic trends using the correlation between the distribution of n-gram noun phrases that represent topics and the link distribution in the citation graph where the nodes are documents containing the phrases [5]. Their approach is based on the intuition that if a phrase is relevant to a topic, the documents containing the phrase have denser connectivity than a random selection of documents. In another example, Mane and Börner denote topics as highly frequent words and words with a sudden increase in usage, a phenomenon called "burst" [6]. Their major sources of these words come from keywords indexed by Institute for Scientific Information (ISI) and MEDLINE's controlled vocabulary, also called MeSH terms. In order to determine the trends of keywords, top 10 most meaningful words were selected by domain experts. The frequency changes of these words over time are used to indicate the trends of each domain.

METHODOLOGY

In this section we describe the topic modeling technique that we use to analyze the research trends in engineering education.

I. Topic Modeling

Topic modeling techniques such as the Latent Dirichlet Allocation model (LDA) [1, 7], aim to identify semantic topics given a text corpus. LDA is a generative probabilistic model of a corpus. It assumes that documents in a corpus are generated as random mixtures over latent topics. Let us assume that there is a corpus with D documents that contain a mixture of multiple topics $\{z_1,...,z_T\}$. LDA specifies the following distribution over words within a document:

$$p(w_i) = \sum_{j=1}^{T} p(w_i \mid z_j) p(z_j)$$

where T is the number of topics. Let $p(w \mid z_j) = \phi^{(j)}$ refer to the multinomial distribution over words for topic z_j and $p(z) = \theta^{(d)}$ be the multinomial distribution over topics for document d. The two sets of parameters, $\varphi^{(j)}$ and $\theta^{(d)}$, indicate which words are important for which topic and which topics are important for a particular document, respectively. Two symmetric Dirichlet distributions with hyperparameters α and β are introduced to the estimation of $\theta^{(d)}$ and $\varphi^{(j)}$, respectively, in order to achieve smoothed topic and word distributions. Those parameters are posterior probabilities that cannot be assessed directly. The values of the hyperparameters depend on number of topics T and vocabulary size. Steyvers suggests that $\alpha = 50/T$ and $\beta = 0.01$ should work well with many different text collections. However, we still need to determine the number of topics T in the corpus. Perplexity is commonly used in language modeling to test the fitness of a text model given training data. A lower perplexity score indicates better generalization performance. Therefore, we can obtain the best approximation of the topic numbers of the data by minimizing the perplexity as: $T = \arg\min_T \{\text{perplexity}(D_{\text{test}} \mid T)\}$.

Following [1, 2], we can evaluate the perplexity on a hold-out test data as:

$$\text{perplexity}(D_{\text{test}} \mid T) = \exp\left(-\frac{\sum_{d=1}^{|D_{\text{test}}|} \log p(w_d \mid T)}{\sum_{d=1}^{|D_{\text{test}}|} N_d}\right)$$

II. Noun Phrase Extension

Frequently occurred noun phrases can also capture the major semantic concepts from a corpus. A noun phrase normally consists of a head noun and optionally a set of modifiers. It is an important grammatical unit of texts in many languages. In natural language processing (NLP), there are two major noun phrase extraction methods, namely static parsing and machine learning. The static parsing method replies on a set of parsing rules pre-defined by linguists. These rules are often described using finite state automation (FSA). However, the effectiveness of this method is strongly dependent on the accuracy and comprehensiveness of the rule set. On the other hand, machine learning methods aim to overcome the drawbacks of static parsing. They rely on various statistical learning techniques to identify important noun phrases by analyze the part-of-speech (POS) tags of texts. Existing machine learning methods include transformation-based method, memory-based method, maximum entropy, hidden markov model, conditional random field, and support vector machine have been reported effective in noun phrase extraction.

III. Keyword Extraction

Keyword extraction is straightforward. It simpliy tokenizes the text to individual words. After removing common stop words (i.e., "a, "the"), you should also remove corpus-specific stop words such as enginnering and education in this particular study. Finally, words are stemmed to their roots (e.g., "studied" to "studi") so as to obtain an accurate vocabulary of the corpus.

SYSTEM DESIGN AND IMPLEMENTATION

Based on the LDA topic modeling technique, we propose a topic trend analysis system. The system consists of 4 modules (see Figure 1). The data controller allows the user to specify the scope of the input corpus by selecting a combination of journals and/or conferences. The context controller asks the user to specify information (title, keyword, or abstract) to be included in the corpus for each publication. The model controller enables to choose the models of extracting topics or concepts in the corpus. It can be either of topic modeling using LDA, noun phrase extraction or keyword extraction. The time controller enables to choose time range to calculate topic trends. It can be either individual years or individual months. Through different selections, a mix of inputs can be obtained giving a view across time and based on different data corpuses. This mechanism ensures that user can apply different lenses on the data.

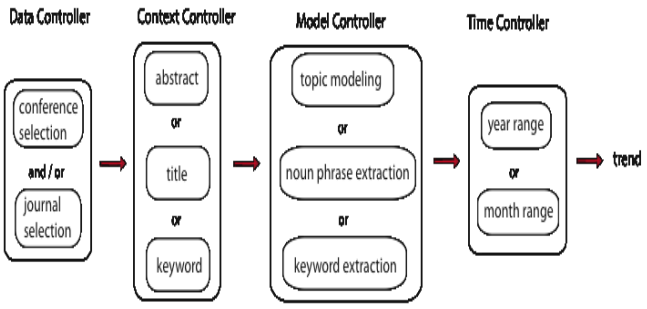

FIGURE 1: SYSTEM OVERVIEW OF TREND ANALYSIS

EXPERIMENT AND DATA ANALYSIS

In this section wed describe data preparation, data analysis, running of experiments on the data, and findings.

I. Date Preparation

We analyze the topic trends on a corpus, consisted of two major journals: Journal of Engineering Education (JEEE) and International Journal of Engineering Education as well as a major conference: Frontiers in Education (FIE). Their publication should cover major researches papers of Engineering Education.

TABLE 1: DATA CORPUS (D is the number of documents, V is the size of vocabulary, W is the total number of words).

Data	D	V	W	Range
JEE, IJEE, FIE	2,645	7,768	203,453	2000-2008

II. LDA Model Estimation

We used an open source LDA package, namely GibbsLDA++[1], for our LDA model estimation. The package is a C++ implementation of LDA using Gibbs sampling technique for parameter estimation and inference. Gibbs sampling is a form of Markov Chain Monte Carlo, which is easy to implement and efficient when extracting a set of topics from a large corpus [3]. We split the original corpus into 90% for training and 10% for testing. We adopt the popular settings for LDA where $\alpha = 50/T$ and $\beta = 0.1$. For Gibbs sampling, we chose to run 1,000 iterations for estimation and 50 iterations for inference. As shown in Figure 2, the LDA model with approximately 60 topics achieved the optimal perplexity score.

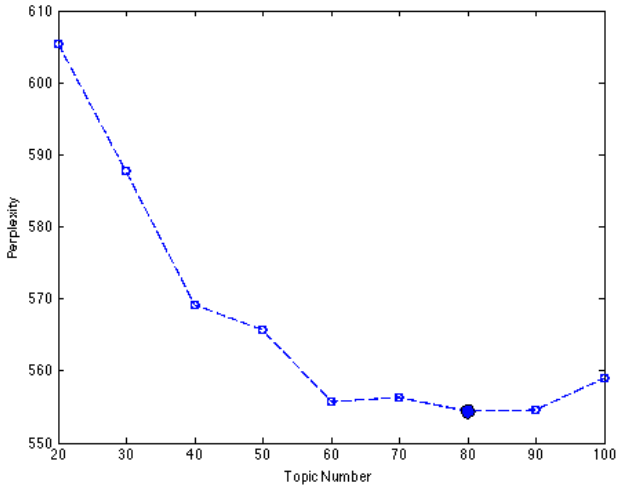

Figure 2: The LDA model with 60 topics achieved the optimal perplexity score

III. Topic Trends

Here, we list the 15 topics out of total with top 10 words in each topic (see Table 2). We analyze the trends of 15 topics using the method introduced in Section. These trends are shown in Figure 3 & 4.

Table 2: Top 15 Topics

Topic #	Top 10 Words in Each Topic
0	Student perform academic study factor significant higher level examination success
1	Design process engineering build idea open support pattern incorporate hand
2	Global intern competition culture university country state unit institution paper
5	Learn instruct base effect strategy cognition evaluation think tradition understand

[1] http://gibbslda.sourceforge.net/

8	School science teacher high student active stem middle career math
12	Device digit application mobile system embed base present logic implement
15	Laboratory lab experiment robot virtual remote control equipment simulation hardware
20	Survey study response result percept relate question rate complete determine
34	Data analysis collect inform analyze quality quantity method generate develop
36	Control simulation electron matlab power circuit paper present operate require
40	Software develop platform paper source potential open provide formal tool
44	Skill community develop technic student profession compete leadership knowledge integrate
49	Method chemic transfer energy numer spreadsheet flow calculate heat fluid
51	Student retent college mentor program success academy freshman increase university
59	Project student design capston require involve senior experiment final manage

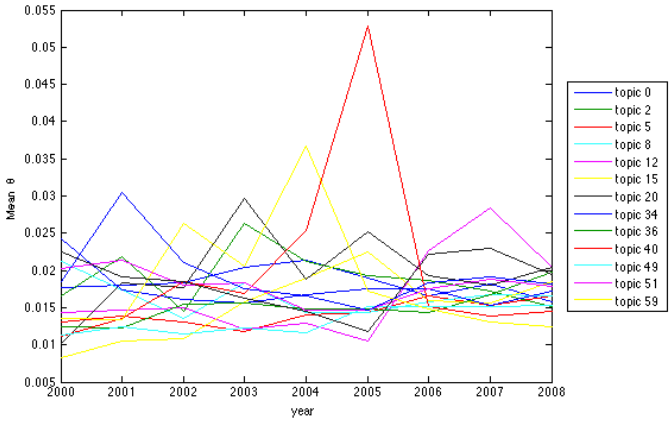

Figure 3: Topic trend of 15 topics between 2000 and 2008

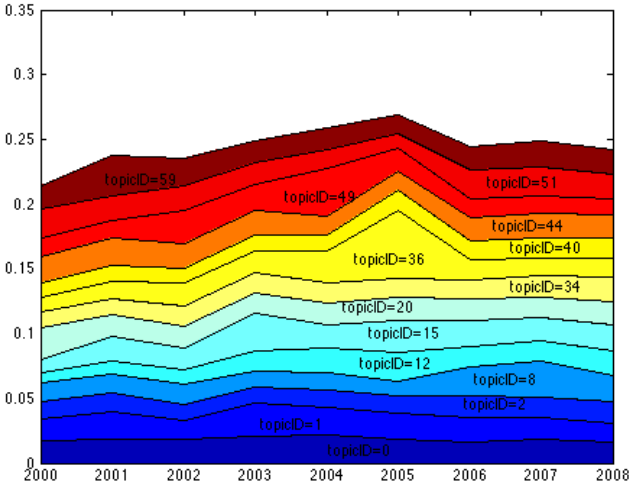

Figure 4: Topic trends of 15 topics between 2000 and 2008

IV. Keyword Trends

We extract top 20 keywords of the entire corpus in Table 3 and analyze their frequency trends over the time. The trends of two representative keywords, "laboratori" and "undergradu", are shown in Figure 5 and Figure 6, respectively.

Table 3: Major Keywords and Their Frequency

Keyword	Frequency		Keyword	Frequency
Learn	633		Project	252
Student	542		Assess	235
Teach	486		Model	234
Base	475		Approach	214
Design	469		Analysis	211
Laboratory	424		Study	211
Chemistry	384		Control	204
Experiment	336		Program	176
Develop	301		Simulate	158
Undergraduate	284		System	158

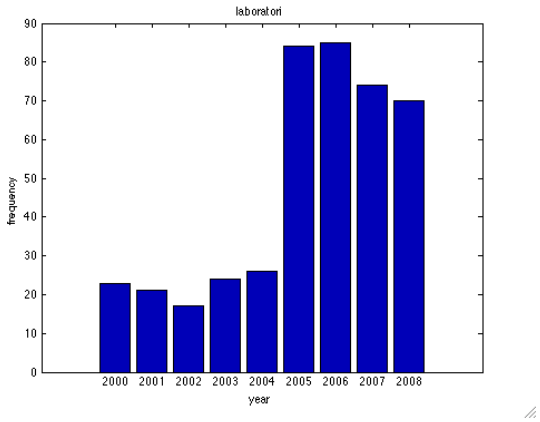

Figure 5: Frequency of Keyword "Laboratories"

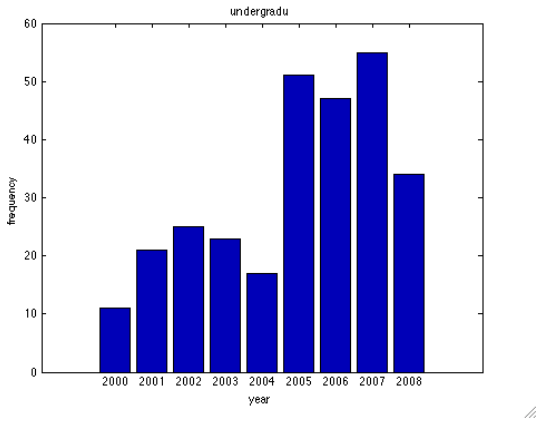

Figure 6: Frequency of Keyword "Undergraduate"

V. Noun Phrase Trends

We extract the top 20 noun phrases of the entire corpus in Table 4 and analyze their frequency trends over the time. The trends of two representative noun phrases, "engineering education" and "a case study", are shown in Figure 7 and Figure 8, respectively.

Table 4: Noun Phrases and Their Frequency

Noun Phrase	Frequency		Noun Phrase	Frequency
students	161		synthesis	40
design	96		evaluation	39
development	93		technology	39
engineering education	92		the role	36
chemistry	64		the impact	32
assessment	56		the use	32
analysis	53		an experiment	31
engineers	48		course	31
a case study	47		research	31
matlab	41		science	31

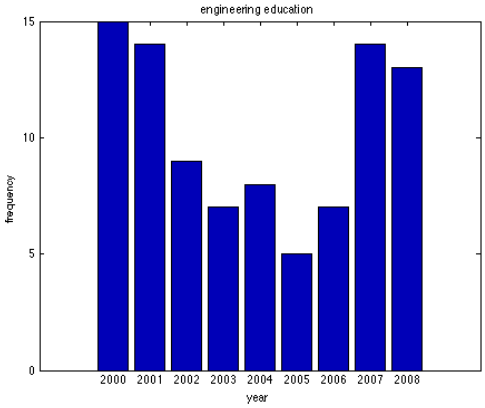

Figure 7: Frequency of Appearance of "Engineering Education"

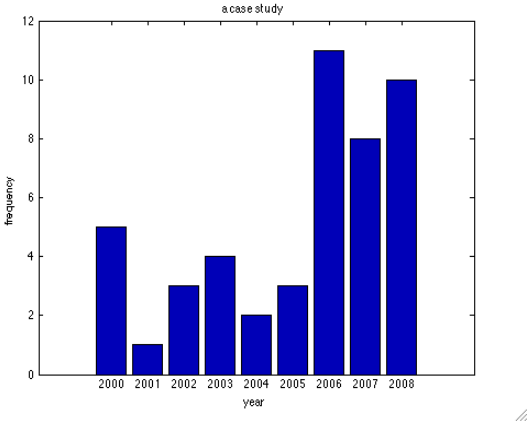

Figure 8: Frequency of Appearance of "A Case Study"

FINDINGS AND DISCUSSION

I. Topic Trends

Overall, the findings from these analyses show that some topics remain constant over time whereas other topics become more popular at certain time periods. For instance, since 2005 the topic of global and international aspects of engineering education has seen a significant spike. This interest can partially be attributed to the discussion of international aspects of educating engineers in the NAE publications (*Engineer of 2020* & *Educating the Engineering of 2020*) as well as the publication of *The World is Flat* by Thomas Friedman which had a significant influence on science and engineering public policy in the United States. The findings from the topic analysis also shed light on several methodological issues that emerged as the primary methods of interest to the community – experiments, case studies and survey-based studies. The results from the analysis also show that certain engineering related software and data analysis tools, such as Matlab, are popular topics for given their use in engineering education and their potential to shape student learning across engineering disciplines. In terms of disciplinary areas, electronic and communications engineering and chemical engineering were found to be common areas addressed by scholars. Efforts such as mentoring and community development were also frequently present in the list of topics. The use of technology in learning was another dominant area of research and several topics (across the analyses) related to technology came up, such as, robotics and mobiles. Not surprisingly, another major topic was design, given the central role of design in engineering practice and engineering learning and cognition. Results from topic modeling also suggested that capstone projects and freshmen projects are an area of interest across the community. Professional skills such as leadership, communication, and teamwork were also part of list of topics that were of interest to a significant number of scholars. Finally, another topic common across all results was assessment.

II. Potential Concerns

One area of potential concern that emerges from the analysis of topics is the lack of any theoretical or analytical keywords. For a growing and maturing field it is essential to develop a body of knowledge, to accumulate knowledge in a meaningful manner [9]-[14]. This body of knowledge can then serve as the basis for productive future research which avoids the pitfalls of earlier efforts. For any academic discipline, particular a social science or interdisciplinary discipline such as engineering education research, it is essential to have strong theoretical or analytical ideas around which a group of scholars can contribute [10]. For instance, no psychological, sociological, or learning sciences theory was present as a keyword. Issues of concern such as student motivation or student identity were also absent from the list of topics. This finding is of significance as it alerts us to a gap between practice and theory and the still greater effort

needed to develop a more cohesive scholarly agenda in the field.

Another area of concern that emerged from the analyses was a disproportionate attention to undergraduate education and a lack of attention to graduate education within the community. Graduate students, in addition to being students of engineering, are also highly involved in both engineering teaching and research. Furthermore, the number of graduate students and their involvement in the engineering and engineering education community is steadily increasing. Therefore, more attention is needed to issues that focus on graduate engineering education. In a related issue, there was no mention of K-12 experiences either, which is also a growing area of interest within engineering education. As the field continues to grow it has to look beyond undergraduate students and steps have to be taken to include graduates and also K-12 students in engineering education and these are potential growth areas. As we further develop our data corpus to make it more inclusive and diverse, we are likely to uncover other areas of interest and of concern to engineering educators.

CONCLUSION

In this paper we describe an approach to assess the growth of a field using topic modeling techniques and apply it to engineering education. By using different approaches to topic modeling we were able to provide a more comprehensive representation of the field than that achievable by other approaches. We combined LDA, noun phrase extraction, and keyword extraction, and all three approaches provided a different lens on the data. We argue that for future work such a combined approach might be the ideal way to understand disciplinary communities and their interests and ideas. We highlight some of the key areas of interest for the community over the past years and identify emerging patterns as well as highlight an area of concern – the lack of theoretical or analytical topics with which the community engages. We also found that interpreting the result occurs best when someone from the disciplinary field looks at the findings.

There major limitation of our work which is the exclusion of non-U.S. venues. Although the journal and conferences in the sample publish international work, their representation is quite limited, therefore skewing the results towards issues that more pertinent to the U.S. In future work we are trying to balance the data by including data from *European Journal of Engineering Education* as well as proceedings from *SEFI* and *REESE*. The goal is to make the dataset as comprehensive and diverse as possible. A secondary concern with the analysis methods adopted here is the frequent occurrence and identification of generic topics such as "students" or "learning." We are cognizant of this issue but also believe that including such topics in the analysis and findings captures a more honest characterization of the field and present a diffuse but real representation of the ideas present in the field.

ACKNOWLEDGMENT

This work was funded through NSF Award#0935124 and #0935090 and a grant from ICTAS at Virginia Tech.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[2] T. L. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, p. 5228, 2004.

[3] T. L. Griffiths and M. Steyvers, "Finding Scientific Topics," in *Proceedings of the National Academy of Science*, 2004, pp. 5228-5235.

[4] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the History of Ideas Using Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, HI, 2008, pp. 363-371.

[5] Y. Jo, C. Lagoze, and C. L. Giles, "Detecting Research Topics Via the Correlation between Graphs and Texts," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Jose, CA, 2007, pp. 370-379.

[6] K. K. Mane and K. Bˆrner, "Mapping Topics and Topic Bursts in Pnas," in *Proceedings of the National Academy of Sciences of the United States of America*, 2004, p. 5287.

[7] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic Author-Topic Models for Information Discovery," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, 2004, pp. 22-25.

[8] X. Wang and A. McCallum, "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia,PA, 2006, pp. 424-433.

[9] Jamieson, L. & Lohmann, J. (2009). "Creating a Culture for Scholarly and Systematic Innovation in Engineering Education". *Phase 1 Report, ASEE*.

[10] Johri, A. "Creating Theoretical Insigths in Engineering Education Research," Journal of Engineering Education, 2010.

[11] Kemnitzer, S. (2008). "The Need for Theory-Based Research in Engineering Education". *Video interview recorded at ASEE 2008*, available online at CASEE video channel on YouTube: http://www.youtube.com/watch?v=nfsh08jHEKs.

[12] Shulman, L. S. (2005). "If not now, when? The timeliness of scholarship of the education of engineers." *Journal of Engineering Education*, pp.11-12.

[13] Streveler, R. & Smith, K. (2006). "Conducting rigorous research in engineering education." *Journal of Engineering Education*, p.103-105.

[14] Watson, K. (2009). "Change in engineering education: where does research fit?" *Journal of Engineering Education*, 98(1):3-4.

AUTHOR INFORMATION

Aditya Johri, Assistant Professor, Department of Engineering Education, Virginia Tech, ajohri@vt.edu

G. Alan Wang, Assistant Professor, Department of Business Information Technology, Virginia Tech, alanwang@vt.edu

Xiaomo Liu, Doctoral Candidate, Department of Computer Science, Virginia Tech, xiaomliu@vt.edu

Krishna Madhavan, Assistant Professor, Department of Engineering Education, Purdue University, cm@purdue.edu