Learning Track Representation and Trends for Conference Analytics

Pengfei Liu[1], Shoaib Jameel[2], King Keung Wu[3], and Helen Meng[1]
[1]*Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong*
[2]*School of Computer Science and Informatics, Cardiff University, United Kingdom*
[3]*Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong*
{pfliu, hmmeng}@se.cuhk.edu.hk, JameelS1@cardiff.ac.uk, kkwu@mae.cuhk.edu.hk

***Abstract*—We propose a new system for facilitating the co-creation of conference tracks through data analytics and human knowledge. The system attempts to learn track representations based on a topic-track matching framework, infer historical track representations by topic evolution paths and investigate the evolution of track representations to figure out track trends. We thus aim to develop a data-driven approach for improving expert-designed track descriptions in future conference organization. One challenge in our work is how to learn track representations from limited publication papers of each year, and another challenge is to how to figure out track trends when track descriptions are not readily available in some years. We present two novel approaches on learning track representation by topic-track matching and analyzing track trends by constructing topic evolution paths, respectively. We also show interesting results on topical leaps and branches from year to year, obtained from papers of INTERSPEECH and ICASSP from 2010 to 2014. These findings should be reflected in the conference tracks to keep them up to date.**

***Keywords*-topic models; conference organization; conference analytics;**

I. Introduction

Academic conferences such as INTERSPEECH, HICSS, CIKM, etc., usually call for papers with a list of tracks to give a high-level description about the scope of an academic conference. In order to further delineate the scope, each track is branched into sub-tracks by domain experts. An example track among the 12 tracks of INTERSPEECH-2014 is given in Table I. Similarly, there is a track named "Electronic Government" from the HICSS conference, which is also further branched with sub-tracks such as "Cybersecurity", "Social Media in Government", and so on. Apart from specifying the scope of the conference, these tracks and sub-tracks help both authors and reviewers respectively in submitting and selecting papers for reviewing. Therefore, they must be designed very carefully in order to facilitate good conference organization.

However, sometimes these tracks may be too broad or ambiguous, or even without any sub-tracks at all. It may be challenging at times to define tracks with the appropriate level of topical granularity and also with clarity. Ambiguity in this definition will cause the authors and reviewers to spend considerable amount of time in searching for the right track when submitting or reviewing the papers. Manual definition of tracks may be laborious, subjective and may not fully capture the state of the development of the field. Therefore, we need to explore automated techniques which can help generate such track and sub-tracks automatically from the text of submitted papers. These auto generated tracks may also augment human-defined tracks to make them more comprehensive. In addition, some topics may gain popularity and others may go out of fashion over time, and such dynamics need to be captured from year to year.

Table I
Descriptions of Track 7 and its Sub-tracks in INTERSPEECH-2014.

7:	**Speech Recognition - Signal Processing, Acoustic Modeling, Robustness, and Adaptation**
7.1	Feature extraction and low-level feature modeling for ASR
7.2	Prosodic features and models
7.3	Robustness against noise, reverberation
7.4	Far field and microphone array speech recognition
7.5	Speaker normalization (e.g., VTLN)
7.6	Deep neural network
7.7	Discriminative acoustic training methods for ASR
7.8	Acoustic model adaptation (speaker, bandwidth, emotion, accent)
7.9	Speaker adaptation; speaker adapted training methods
7.10	Pronunciation variants and modeling for speech recognition
7.11	Acoustic confidence measures
7.12	Multimodal aspects (e.g., AV speech recognition)
7.13	Cross-lingual and multilingual aspects, non native accents
7.14	Acoustic modeling for conversational speech (dialog, interaction)

In this paper, we aim to develop a system for the co-creation of conference tracks through analyzing publication papers and integrating human knowledge. We wish to address the following issues:

(I) Can we find an intuitive, data-driven approach to visualize each track instead of a few key words?
(II) Can we analyze the popularity of a track and its trends over the five years?
(III) Can we find the historical footprint of a sub-track, e.g., "*deep neural network*"?

First, we propose a topic-track matching task between the latent topics inferred from paper abstracts based on a topic model named latent Dirichlet allocation (LDA) and the expert-designed tracks. Topic models have proven to be very effective in finding out the latent dimensions of text data, thereby bringing out the set of thematic words which we call a topic. These thematic words can be used to describe a track in our task. In addition, large scale analysis using these models has been done and proven to be effective as well.

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.211

The reason we use LDA for our task is that it is effective and efficient to apply on large text collections [1], [2] and it has been applied successfully on analyzing scientific documents [3], [4].

Second, we represent each human-defined track with its matching topics, visualize the high-probability keywords of each topic with word cloud and figure out the most relevant papers of each topic. This will help highlight some representative terms in each track, which later can be used in refining tracks and its sub-tracks or introducing a completely new track.We aim to improve track descriptions based on a data-driven approach where research papers themselves guide the organizers.

Third, we analyze the trends of each track by investigating the evolution of track representations year by year. Historical track representations are inferred by constructing the topic evolutionary paths backwards year by year for each matching topic of a track. A topic evolutionary path also enables us to observe trends of a certain topic on its appearance, being popular and being branched etc. Conducting such analytics on conference publications related to a particular research area will help us better understand its developments and trends, and help research newcomers to undertake a popular topic in their research area.

There are a few challenges in this research: (1) We have limited data to work with. Topic models are statistical models that conduct co-occurrence analysis on text data. The more data, the better are the captured latent topics. In order to address this challenge, we adopted hyperparameter (hyper-prior) sampling technique to tune the hyperparameters of the model based on the data itself. This is because the hyperparameters can help get rid of some general words from the latent topics [5]. (2) Selecting the appropriate number of topics is also very challenging. Using an arbitrary number of topics will not help us achieve our goal satisfactorily. In addition, tuning based approaches such as the one used in [6], may not lead to satisfactory results because the number of papers each year is relatively small. Therefore, we used the non-parametric Hierarchical Dirichlet Process (HDP) [7] model to find out the number of topics automatically based on data characteristics. (3) We do not have historical track descriptions since most of them can no longer be found online, which makes it hard to analyze track trends over time. We thus adopted a novel data-driven approach to represent each available track (e.g., the tracks from INTERESPEECH-2014) with a set of latent topics, inferred historical track representations using topic evolutionary paths and analyzed track trends by investigating the evolution of track representations.

In the remainder of this paper, after discussing related work in Section II, we present the system blueprint and our approach in Section III. Section IV shows the corpus, experimental settings, and the results on track representation and trends; and Section V discusses the advantages and disadvantages of our approach. We conclude the paper and propose future work in Section VI.

II. Related Work

Probabilistic topic models [8], [9] such as latent Dirichlet allocation (LDA) [10] are statistical models that find patterns of words or underlying latent topics from a large collection of documents, which have been widely used in the past to study academic conferences [3], [11], [12], [9], [13].

In LDA, documents are represented as random mixtures over latent topics and each topic is characterized by a distribution over words [10]. LDA ignores word order information but considers each document as a bag of words, and assumes a flat one-level topic structure. Moving beyond the bag-of-words assumption, [14] presented a bigram topic model which extended LDA by incorporating word order information similar to a hierarchical Dirichlet bigram language model and showed better predictive accuracy than LDA.

To learn topic hierarchy from data, [11] presented a hierarchical latent Dirichlet allocation (HLDA) model which combines a nonparametric prior by a nested Chinese restaurant process with a likelihood based on a hierarchical variant of latent Dirichlet allocation. They applied HLDA on 1717 paper abstracts of NIPS from 1987 to 1999 and demonstrated a 3-level topic hierarchy, with the first level capturing the function words, the second level separating the words pertaining to *neuroscience* and *machine learning* fields, and the third level delineating several important subtopics within the two fields.

To model correlation between topics and overcome the limitation of LDA stemmed from the independence assumptions implicit in the Dirichlet distribution on the topic proportions, [15] introduced the correlated topic model (CTM) where the topic proportions are drawn from a logistic normal distribution to model correlations by the covariance matrix, rather than a Dirichlet distribution with strong independence assumption among topics. CTM gives a more realistic model of latent topic structure where the presence of one latent topic may be correlated with the presence of another, as illustrated in the topic graph learned from 16,351 OCR articles from Science by [15].

With the rapid increase of academic papers, there is a great demand for analyzing historical trends of a research field from these papers. To capture the dynamic evolution of topics in a sequentially organized corpus of documents, [16] presented a dynamic topic model (DTM) which chains together topics and topic proportion distributions over time by Gaussian distributions, i.e., by using Gaussian time series on the natural parameters of the multinomial topics and logistic normal topic proportion models. They conducted experiments on a paper corpus from over 100 years of OCR'ed articles from the journal *Science*. DTM offers new ways of browsing large, unstructured document collections. However, DTM assumes a fixed number of flat topics (i.e., no modeling of hierarchical structure among topics) evolved using Gaussian time series, and does not explicitly model the rise and fall in popularity of a topic or in the use of specific terms.

Similarly, [17] introduced a non-Markov continuous-time model named topics over time (TOT) for capturing topical trends. Compared with LDA, TOT adds an additional component of *time stamp* generated by a per-topic Beta distribution. TOT assumes topics and their meaning are constant over time and thus captures changes in topic co-occurrence instead of changes in the word distribution of each topic. [18] developed a dynamic hierarchical Dirichlet process model named infinite dynamic topic models (iDTM), which allows for unbounded number of topics. iDTM adopted the recurrent Chinese restaurant franchise (RCRF) process to model evolution of topic popularity, topic word distributions and the number of topics over time.

In addition, there are some other recent related applications. [2] used topic modeling approach on research papers to study how popularity of topics change over time. In [19], the authors studied a comparison based approach of different rankings of topics to discover topics with persistent, withering and booming establishment in a scientific field using the LDA model. [20] studied the ideas and the dynamics in a research community, but did not use topic models to address their task.

Most relevant to our work are the work by [3] and the work by [4]. [3] applied LDA on the whole paper abstracts in PNAS from 1991 to 2001, identified *hot* and *cold* topics using topic dynamics and highlighted the semantic content in abstracts by tagging each word with its topic assignment; while [4] used LDA to analyze historical trends in the field of *computational linguistics* from 1978 to 2006. Different with the both works, our work in this paper mainly focuses on *learning track representations* from conference papers, *investigating track trends* by analyzing the matching latent topics of each track, and aiming to improve future conference organization with a data-driven approach by topic modeling techniques.

III. Approach

This section describes the system blueprint and our approach for analyzing conference papers of INTERSPEECH and ICASSP from 2000 to 2014. First, we applied the LDA model on the paper abstracts of each year to infer latent topics and used HDP to find the number of topics automatically. Second, for each track in INTERSPEECH-2014, we found out its matching latent topics by a similarity-based matching procedure. Third, for each latent topic inferred from INTERSPEECH-2014, we connect it with its most similar topic in 2013, and then build a *topic evolution path* by connecting topics from neighboring years year by year. Finally, we analyzed track trends by investigating each topic evolution path of the matching topics of each track in INTERSPEECH-2014.

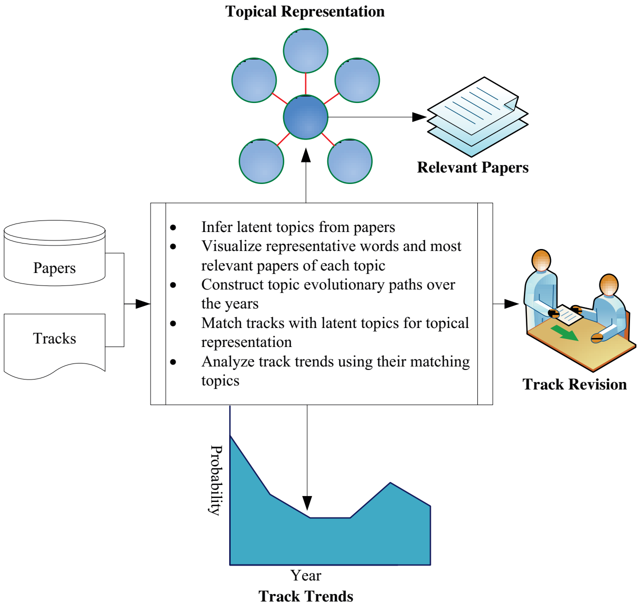
Topical Representation
Relevant Papers
Papers
Tracks
Infer latent topics from papers
Visualize representative words and most relevant papers of each topic
Construct topic evolutionary paths over the years
Match tracks with latent topics for topical representation
Analyze track trends using their matching topics
Track Revision
Probability
Year
Track Trends

Figure 1. The System Blueprint.

A. System Blueprint

In order to serve research communities better on academic conference organization, we propose to develop a *smart* system that can *understand* research trends of a field, integrate human knowledge (e.g., expert-designed conference tracks) and *figure out* how to organize a conference to reflect the development of a research area. The system is outlined in Figure 1, whose inputs include papers and tracks, and outputs include a topical representation together with the most relevant papers for each track. The system also analyzes track trends through matching topics and suggests possible wording revisions for the tracks through data analytics and human knowledge.

B. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic topic model for collections of discrete data such as text corpora. The aim of LDA is to represent the meaning of each document as a probability distribution over a set of latent topics, which in simple terms means that documents exhibit multiple topics. Each latent topic is represented by a probability distribution over words in the vocabulary. It assumes a generative process, in which each word in a document is generated by sampling a topic and then sampling a word. LDA assigns each word in the document collection to one of the latent topics (initially at random), and uses this assignment to estimate both the probability distribution over topics associated with each document and the probability

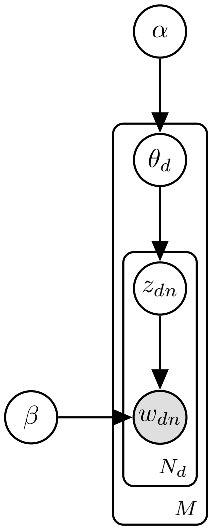

Figure 2. Graphical representation in plate diagram of the LDA model. Plates signify repetition of variables and unshaded circles represent variables in the probability model. Shaded variable in the model depicts observed variable.

distribution over words associated with each topic. These probability distributions are then used to improve the topic assignments of the words, and the whole process is repeated until convergence. We depict the graphical representation of the LDA model in Figure 2 to illustrate the relationships between different variables in LDA.

In the graphical model, θ_d for each document d is the topic proportion for that document. This variable tells us about the importance of a latent topic to a document, and it has a prior α which comes from a Dirichlet distribution. The reason to choose a Dirichlet distribution is that the topics are sampled from a multinomial distribution whose conjugate prior is a Dirichlet distribution. This choice makes the posterior distribution also a Dirichlet distribution that leads to simple models and easier computing. An observed word in the document w_{dn} is generated by a latent topic index variable z_{dn}. This index variable is the topic assignment variable for that word. The variable β contains the word-topic distribution. This is a matrix of the number of topics multiplied by the size of the vocabulary, where vocabulary contains all the unique words in the text collection. The input to the model is the term document matrix and the LDA model outputs two distributions which are document-topic distribution and word-topic distribution. In this model, we have to specify the number of topics a priori, which means that the dimensionality of the topic space is fixed and the assumption is that the user is aware of the number of topics in advance.

Following the notations in [10], given the hyper-parameters α and β, LDA defines the probability of a corpus D with M documents, as illustrated in Equation (1), where θ_d is a topic mixture for document d, w_{dn} is the nth word from document d, and z_{dn} is the latent topic assignment for the word w_{dn} given θ_d.

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d \quad (1)$$

In the generative process of LDA, each document is generated by first sampling a document-specific topic proportion θ_d from a Dirichlet distribution, and then drawing each word from a topic-specific multinomial distribution $p(w_{dn}|z_{dn},\beta)$. The model generates a low-dimensional representation of data, consisting of a word distribution of $P(w|z)$, which states the probability of a word w belonging to a topic z and a topic distribution in a document $P(z|d)$, which specifies the mixture of topics in a document d. Our interest is on $P(w|z)$ as we will match the words in topics with the words in expert-designed tracks of an academic conference. The LDA model can be estimated by several algorithms, such as the variational Bayes algorithm by Blei et al. [10], the expectation propagation algorithm by Minka et al. [21] and the collapsed Gibbs sampling algorithm by Griffiths and Steyvers in [3] and so on, which are compared in [22] and [23].

C. Hierarchical Dirichlet Process

The Hierarchical Dirichlet Process (HDP) model is a non-parametric Bayesian model [7], [24], [25], [9], which can be used to automatically find out the number of latent topics based on data characteristics along with the topic discovery task as usual. HDP extends its parametric part of LDA by adding an estimator for the number of topics and associated parameters, which is thus able to find the number of topics automatically from the dataset. Although in [7] the authors presented the HDP model in general, it can be adapted for topic modeling as done in many works such as [26]. We could have considered using cross validation approach to automatically find out the number of latent topics using the LDA model, but in our case the dataset consists of a small set of papers in each year which might not be practical in finding a desirable number of latent topics. Therefore, we used the HDP model in our task. The HDP model can vary its complexity based on the sample size, which means that the parameters in the model can grow and shrink based on the data characteristics. Incorporating the hierarchical nature in the model introduces a mixed-membership property in which sharing among the clusters can occur. This special sharing property brings out a variety of relationships among the clusters in the topic space.

Given a collection of the text documents, HDP is characterized by a set of random probability measures G_d for each document d in the collection. A global random probability

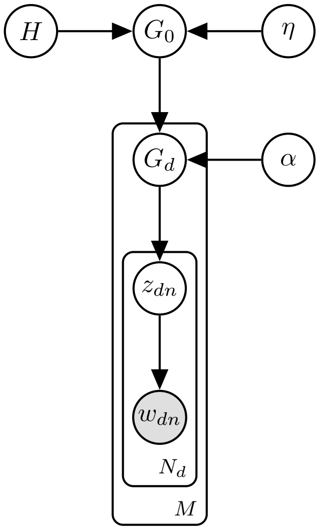

Figure 3. Graphical representation in plate diagram of the HDP model.

measure G_0 which is drawn from a Dirichlet process (DP) with a base probability measure H. The global measure G_0 selects all the possible topics from the base measure H, and then each G_d draws the topics necessary for the document d from G_0. The concentration parameter to this base measure is denoted as η which is used to regulate the variance around the distributions. Similarly, α is the concentration parameter to the distribution associated with G_d. Note that the role played by the α values in the HDP and the LDA models are different, but variables such as w_{dn} and z_{dn} have the same meaning here as in the LDA model. We depict the graphical representation of the model in Figure 3.

From the above description, it is clear that the HDP model describes a distribution over distributions as a hierarchy. There are different metaphors which is used to describe the HDP model. These metaphors are mainly used to sample data from the HDP. Metaphors such as Chinese Restaurant Franchise (CRF), Polya Urn, etc are popularly used. The model that we used in our experiments uses the Chinese Restaurant Franchise based sampler. In this scheme, there are two levels of Chinese Restaurant Processes (CRP) which is again a metaphor to describe the Dirichlet process. The HDP model makes use of CRF metaphor to generate samples from the posterior distribution given the observations. In order to describe the sharing among the groups, the notion of "franchise" has been introduced that serves the same set of dishes globally. When applied to text data, each restaurant corresponds to a document. Each customer corresponds to a word. Each dish corresponds to a latent topic. A customer sits at a table, one dish is ordered for that table and all subsequent customers who sit at that table share that dish. The dishes are sampled from the base distribution which corresponds to discrete topic distributions. Multiple tables in multiple restaurants can serve the same dish. A table can be regarded as the topic assignment of the words in documents. Inquisitive readers are requested to consult [7] for more technical details about the model.

D. Topic-Track Matching

Conceptually, we assume a *one-to-many* relationship between human-defined track and automatically generated topic, and we define a match as K tuples between a topic and its *top-K* similar tracks based on F-score. We tackle the *topic-track matching* problem under an information retrieval framework, with each latent topic as a *query* and each track as a *document*.[1] The latent topics are obtained by applying LDA on the paper abstracts and we represent each topic by choosing its most probable 200 words based on the decreasing probability of each word. We have empirically found out that the most probable 200 words generally cover approximately 80% of the probability space of the words in each topic in our datasets.

The flowchart of topic-track matching is illustrated in Figure 4. We first applied the same pre-processing step to both *Conference Tracks* and *Paper Abstracts*. Then, we applied LDA to get the latent topics with a list of *top words* in descending order of probability $P(w|z)$, which are queries for retrieving the tracks represented with a set of *key words*. For each *query* (topic), we match it with the *document* (track) which has the highest F-score obtained by calculating their overlapping words. Each *document* typically consists of 20-50 words pre-processed from the corresponding track description. Similar to the F-score measure used in *Text::Similarity*[2] for pair-wise similarity of files or strings, we calculated F-score by first counting the number of matching words between the key words of a track (W_k) and the top words (W_t) of a topic, and then computing *Precision*, *Recall* and *F-score* using Equation (2), (3) and (4), respectively.

Note that F-score is essentially a variant of the *Jaccard coefficient* [27] of two sets of elements, as shown in formula (4) and (5). The F-score and the Jaccard coefficient output the same matching results because their numerators are the same while the denominators are slightly different normalization constants.

$$\text{Precision} = \frac{|W_k \cap W_t|}{|W_k|} \tag{2}$$

$$\text{Recall} = \frac{|W_k \cap W_t|}{|W_t|} \tag{3}$$

$$F\text{-score} = \frac{|W_k \cap W_t|}{\frac{|W_k|+|W_t|}{2}} \tag{4}$$

$$\text{Jaccard} = \frac{|W_k \cap W_t|}{|W_k \cup W_t|} \tag{5}$$

[1] We also tried to apply LDA on the tracks directly to infer their top covered topics, which are however not distinguishable among tracks.

[2] https://metacpan.org/pod/Text::Similarity

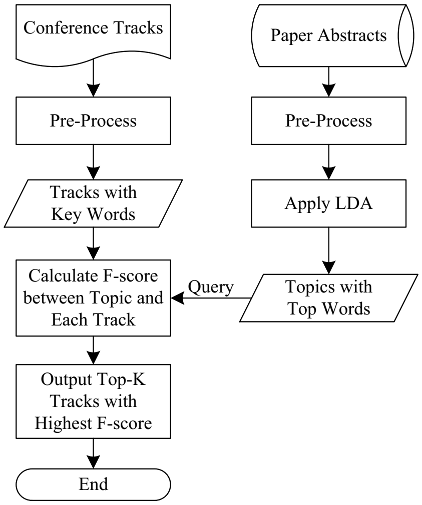

Figure 4. The Flowchart of Topic-Track Matching.

E. Track Representation

Tracks are typically defined by domain experts when organizing an academic conference, which consist of a set of key words to describe a research area, e.g., "acoustic modeling" in the speech recognition track. These tracks are useful for the organization of a technical conference where authors can submit their papers to their preferred tracks, reviewers can choose papers to review from their preferred tracks, and readers can search for papers by filtering tracks. Sometimes these tracks are not clearly defined, or do not even exist in the submission system which leads to plenty of issues where conference organizers have to respond to individual queries of the authors.

Alternatively, we learn track representations from a collection of papers automatically to save time-consuming human efforts. Moreover, they are more likely to be up to date with the development of a field than the static track descriptions of a few words or phrases. In addition, they provide an alternative perspective to understand each expert-designed track, by showing *most probable* words, visualizing with word cloud, pointing out relevant papers, and thus could assist track revisions for future conference organization.

In this work, we define *track representation* as a set of latent topics matched with a track using our proposed topic-track matching framework, together with their most probable words, the graphical visualization of each matching topic (e.g., word cloud), as well as the most relevant papers for each topic.

F. Track Trends

Track trends are defined as the *evolution of track representations* over the years. The challenge is that track descriptions are only available in some years, e.g., INTERSPEECH-2014, which makes it hard to learn track representations for some years without available track descriptions. We tackle this challenge by constructing topic evolutionary paths for the matching topics of each track in INTERSPEECH-2014. Thus, we are able to investigate the evolution of track representations over the years. We build connections for topics from neighboring years based on pairwise topic similarity, computed by both the *Kullback-Leibler* (KL) divergence and the *Hellinger* (HL) distance. KL divergence measures the information lost when approximating a probability distribution P with Q, while HL distance computes the squared distance between two probabilities, as defined in Equation (6) and (7) respectively for the discrete case:

$$KL(P||Q) = \sum_{i=1}^{K} P(i) \log \frac{P(i)}{Q(i)} \quad (6)$$

$$HL(P||Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{K} \left(\sqrt{P(i)} - \sqrt{Q(i)}\right)^2} \quad (7)$$

IV. Experiments

This section presents our corpus, experimental settings and results for the two main tasks: (1) learning topical track representations for INTERSPEECH-2014 by topic-track matching; (2) analyzing track trends by building topic evolutionary paths for their matching topics.

A. Corpus

Our corpus consists of two datasets: INTERSPEECH and ICASSP, which contain the paper abstracts from the conferences of INTERSPEECH and ICASSP for 15 years from 2000 to 2014, respectively. The basic corpus statistics are shown in Table II.

Table II
Basic Statistics of the Corpus.

	INTERSPEECH	ICASSP
No. of Papers	11,130	19,021
Total Words	1,493,823	2,356,496
Vocabulary Size	25,730	30,739

B. Experimental Settings

Pre-processing: For pre-processing, we kept only content words and stemmed each word by morphology (i.e., computing the base form of English words by removing inflections such as noun plurals, pronoun case and verb endings). We lowercased all words and removed stop words which are in the default stop word list provided by MALLET.

Hyperparameters of α and β: We applied the LDA model on the pre-processed dataset, with the hyper-parameters optimized using the hyper-parameter sampling algorithm implemented in MALLET. An advantage of adopting this technique is that the parameters are automatically tuned based on the data characteristic itself.

Number of Topics: We applied the HDP model on each year's papers to find the number of topics automatically, as shown in Table III.

Number of Most Probable Words: We chose the most probable 200 words to represent each topic. This number is chosen empirically, which covers approximately 80% of the probability space of each topic.

Number of Matching Topics: In the step of topic-track matching, we empirically matched each topic with its *top two* most similar tracks, based on the largest two F-scores. Top-one matching means no overlapping topic among tracks while it makes more sense to have fine-grained overlapping topics in different tracks. For example, "deep neural networks" may exist for both the *speech recognition* track and the *speech synthesis* track.

C. Track Representation

INTERSPEECH-2014 has defined 12 tracks with a few sentences and some key words. Beyond that, we are interested in learning a topical representation for each track from the papers, which may give us an intuitive interpretation of the corresponding track. We applied LDA on the papers in INTERSPEECH-2014 and found the matching topics for each track using our topic-track matching framework described in Section III-D.

As an example, Table IV presents the 10 matched topics of Track 7 (See Table I) in INTERSPEECH-2014, and the top 15 representative words of each topic with their probabilities in descending order. We may use these words to revise the track descriptions, e.g., *intelligibility, recurrent, lstm* to reflect the keywords changes in latest papers.

We further visualized each matching topic with *word cloud*. Figure 5 shows a word cloud of *Topic 31* (the fifth topic in Table IV) by the tool Wordle[3] and the matching words with Track 7, where words with larger fonts have higher probabilities in the topic. The layout and color are randomly set for visualization. The visualization suggests that "deep" and "recurrent" neural network has become a popular research topic for Track 7.

In addition, we show the two papers that have highest topic proportions of Topic 31:

(1) Hasim Sak, Oriol Vinyals, Georg Heigold, Andrew Senior, Erik McDermott, Rajat Monga, and Mark Mao. "Sequence Discriminative Distributed Training of Long Short-Term Memory Recurrent Neural Networks." In *Proceedings of INTERSPEECH*, 2014.

(2) Hasim Sak, Andrew Senior, and Franoise Beaufays. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." In *Proceedings of INTERSPEECH*, 2014.

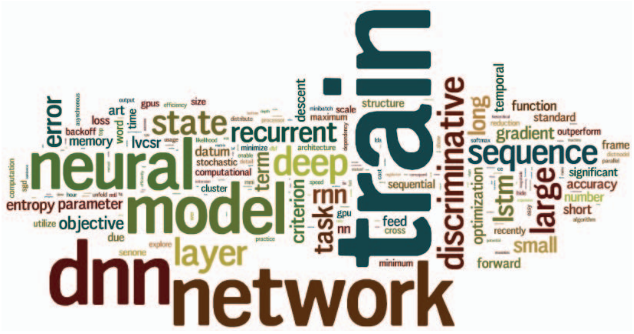

Figure 5. Word Cloud of Topic 31 related with *Neural Network*: it matches with Track 7 with the F-score of 0.186 and the matching words are: *cross, deep, discriminative, model, network, neural, train*.

They both focus on *long short-term memory recurrent neural networks*. Conference organizers may check these papers further to determine whether there is an emerging trend on LSTM for speech recognition and decide how to reflect this trend in the next conference. As a reference, the abstract of the first paper is shown below:

> We recently showed that Long Short-Term Memory (LSTM) recurrent neural networks (RNNs) outperform state-of-the-art deep neural networks (DNNs) for large scale acoustic modeling where the models were trained with the cross-entropy (CE) criterion. It has also been shown that sequence discriminative training of DNNs initially trained with the CE criterion gives significant improvements. In this paper, we investigate sequence discriminative training of LSTM RNNs in a large scale acoustic modeling task. We train the models in a distributed manner using asynchronous stochastic gradient descent optimization technique. We compare two sequence discriminative criteria maximum mutual information and state-level minimum Bayes risk, and we investigate a number of variations of the basic training strategy to better understand issues raised by both the sequential model, and the objective function. We obtain significant gains over the CE trained LSTM RNN model using sequence discriminative training techniques.

D. Track Trends

We study the temporal trends of research tracks through analyzing the dynamical evolution of their matching topics, i.e., the evolution of track representations. Specifically, for each matching topic of a track, we first built its topic evolutionary chain over the past years (See Figures 6 and 7), and then calculated the mean topic probability of each topic in the chain (See Figure 7). As some evolutionary paths may join on certain years, we can then see topic branchings from these join points. For topics with the same original topic, we also compute their accumulated topic probabilities and plot the changes, with an example shown in Figure 8.

Take *Track 7* in INTERSPEECH-2014 as an example, we found its 10 matching topics, shown in Table IV, in the step of topic-track matching. For each matching topic, we found

[3] http://www.wordle.net/

Table III
NUMBER OF TOPICS INFERRED BY HDP FOR THE CONFERENCE PAPERS OF INTERSPEECH (INT.) AND ICASSP (ICA.) FROM 2000 TO 2014.

	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14
INT.	37	40	37	41	38	37	44	40	37	37	32	38	44	43	38
ICA.	39	42	32	36	36	33	34	34	34	30	41	38	42	38	35

Table IV
THE TOP 15 WORDS OF EACH MATCHED TOPIC OF TRACK 7. THE NUMBER OF TOPICS WAS SET TO 38, DETERMINED BY HDP. THE TABLE HEADER SHOWS THE MATCHING F-SCORE.

0.267	0.259	0.202	0.186	0.162	0.159	0.144	0.126	0.09	0.057
dnn	speech	accent	train	speaker	model	method	frequency	listener	vocal
network	noise	english	dnn	adaptation	hmm	signal	feature	native	tract
neural	enhancement	french	network	space	mixture	propose	representation	vowel	vowel
deep	signal	native	neural	datum	parameter	estimate	spectrum	study	formant
feature	noisy	asr	model	diarization	method	filter	time	consonant	acoustic
layer	method	german	deep	experiment	hide	frequency	mfcc	result	space
learn	algorithm	divergence	layer	vector	cluster	time	filter	english	shape
train	propose	effect	sequence	map	gaussian	source	coefficient	phonological	frequency
gmm	phase	phonetic	state	individual	conventional	obtain	signal	learner	method
acoustic	result	speaker	discriminative	corpus	utterance	estimation	cepstral	language	function
bottleneck	snr	speech	large	adapt	markov	array	spectral	target	control
hide	intelligibility	bilingual	recurrent	identity	component	base	band	show	characteristic
convolutional	process	lexical	lstm	similarity	acoustic	localization	mel	contrast	talker
recognition	clean	context	rnn	match	predict	reconstruction	domain	present	intelligibility
architecture	enhance	kullback	task	unsupervise	base	field	bandwidth	speaker	area

out its most similar historical topic in 2013 which has the least KL divergence or the smallest Hellinger distance, and thus built the topic connection from 2013 to 2014. We then applied the same connection procedure to other neighboring years and obtained the final topic evolution paths *backwards* to find the most similar historical topic year by year. Two examples of the 5-year topic connection chains from 2010 to 2014 on INTERSPEECH and ICASSP are shown in Figure 6, which reveal the similar trends of *deep neural networks* from emergence to being popular. In our experiments, the KL divergence and the Hellinger distance output the same results for the two topic evolutionary chains.

For all the matching topics of Track 7, we constructed their topic evolutionary paths and found that some paths joined on certain years. Figure 7 shows the three topic evolutionary paths in red *(a)*, blue *(b)* and green *(c)* for three matching topics of Track 7 from 2010 to 2014. We can see that the topic in 2011 was branched into two topics in 2012 and the upper topic in 2013 was branched into two topics in 2014. Interestingly, we found that the topic related with *deep neural network* emerged in 2012 and a new branch related with *recurrent neural network* became popular in 2014.

Besides, for each topic in Figure 7, we calculated its mean topic probability (on top of each block of words) over all papers year by year. As those topics are evolved from the same topic in 2010, we accumulate the probabilities of all the connected topics for each year and plot the probability changes in Figure 8, which shows *deep neural network* are becoming more and more popular from 2012 to 2014.

V. DISCUSSION

We take a data-driven approach for the co-creation of conference tracks through data analytics and integrating human knowledge. We investigate a novel LDA-based *topic-track matching* framework, which provides an automatic way to infer latent topics from large corpus of conference papers and match them with expert-designed tracks for learning track representations and capturing track trends. By matching a track with a list of latent topics, we are not only able to represent each track with its matching topics but also able to capture track trends when applying the matching on

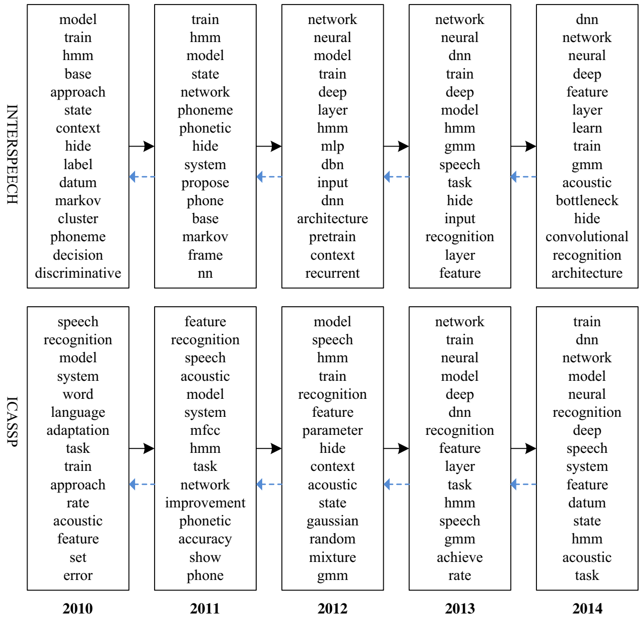
INTERSPEECH
model train hmm base approach state context hide label datum markov cluster phoneme decision discriminative
train hmm model state network phoneme phonetic hide system propose phone base markov frame nn
network neural model train deep layer hmm mlp dbn input dnn architecture pretrain context recurrent
network neural dnn train deep model hmm gmm speech task hide input recognition layer feature
dnn network neural deep feature layer learn train gmm acoustic bottleneck hide convolutional recognition architecture
ICASSP
speech recognition model system word language adaptation task train approach rate acoustic feature set error
feature recognition speech acoustic model system mfcc hmm task network improvement phonetic accuracy show phone
model speech hmm train recognition feature parameter hide context acoustic state gaussian random mixture gmm
network train neural model deep dnn recognition feature layer task hmm speech gmm achieve rate
train dnn network model neural recognition deep speech system feature datum state hmm acoustic task
2010
2011
2012
2013
2014

Figure 6. Two examples of the topic connection chains on INTERSPEECH (top) and ICASSP (bottom) from 2010 to 2014. Consecutive topics are most similar with each other and the chains are constructed backwards year by year as indicated by the blue dashed lines.

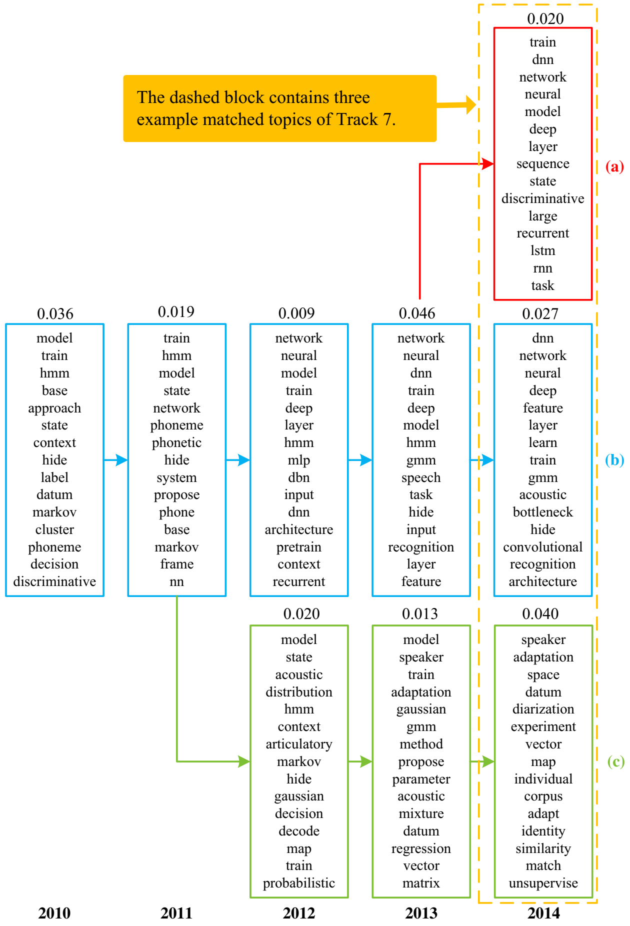

Figure 7. An example trends for the three matched topics of Track 7 in INTERSPEECH-2014. The number on the top of each block of words is the mean topic probability over all papers each year.

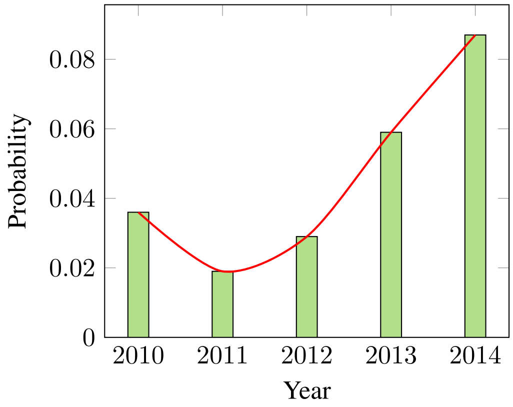

Figure 8. Accumulated topic probabilities for the example trends in Figure 7.

the conference papers over the years.

We chose topic models for our task primarily because these models have proven to be very effective in finding out the latent dimensions of text data, thereby bringing out the set of thematic words which can be used to describe a track in our task. We used LDA for our task because LDA is simple to apply on large text collections and it has shown to be effective on many different tasks, such as analyzing scientific documents.

Track representation represents each track with a set of latent topics from a structured perspective. It provides a topical decomposition of a track, and thus enables us to visualize the matching topics and find out the most relevant papers. We build topic evolutionary paths by connecting the most similar topics from neighboring years backwards from year to year. These paths enable us to infer historical track representations and then analyze *track trends* over years.

However, our approach has some limitations. First, it needs the detailed track descriptions, which are not readily available in some conferences and requires domain experts to define them manually. Second, it would be better to introduce a hierarchical topic representation for each track instead of the flat topic structure learned by LDA.

VI. Conclusions and Future Work

In this paper, we compiled two datasets from conference papers of INTERSPEECH and ICASSP from 2000 to 2014 for the speech community, and proposed a conference analytics system in order to serve research communities better on conference organizations. As research topics are evolving over the years, we inferred latent topics from the two datasets using LDA year by year and built the topic evolutionary paths by connecting the similar topics of consecutive years. We visualized the topic evolutionary paths, word clouds and most relevant papers for topics of interest. Besides, we matched latent topics with expert-designed conference tracks to find a better topical track representation rather than just a few words and investigated track trends through its matching topics. We conclude that our system is capable of finding matching topics of a track (topic representation of a track) and capturing its trends over years (dynamic property of a track). We also share our system open source with the research community.

An interesting future direction is how to assist conference organizers to find reviewers through data analytics. We may extend our framework of *topic-track matching* to a similar idea of *paper-reviewer matching* to pick reviewers by analyzing the similarity between a reviewer's publications with a submitted paper.

Acknowledgments

We are grateful to the anonymous reviewers for their insightful comments and suggestions to improve the paper. This research is affiliated with the Stanley Ho Big Data Decision Analytics Research Centre of The Chinese University of Hong Kong.

References

[1] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed Gibbs sampling for latent Dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 569–577, ACM, 2008.

[2] E. Yan, "Research dynamics: Measuring the continuity and popularity of research topics," *Journal of Informetrics*, vol. 8, no. 1, pp. 98–110, 2014.

[3] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[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*, pp. 363–371, ACL, 2008.

[5] H. M. Wallach, D. M. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," in *Advances in Neural Information Processing Systems*, pp. 1973–1981, 2009.

[6] S. Jameel, W. Lam, and L. Bing, "Supervised topic models with word order structure for document classification and retrieval learning," *Information Retrieval Journal*, pp. 1–48, 2015.

[7] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, 2006.

[8] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of Latent Semantic Analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[9] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

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

[11] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," *Advances in Neural Information Processing Systems*, vol. 16, p. 17, 2004.

[12] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272, Association for Computational Linguistics, 2011.

[13] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine learning*, vol. 95, no. 3, pp. 423–469, 2014.

[14] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 977–984, ACM, 2006.

[15] D. M. Blei and J. D. Lafferty, "Correlated topic models," *Advances in Neural Information Processing Systems*, vol. 18, p. 147, 2006.

[16] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, ACM, 2006.

[17] 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*, pp. 424–433, ACM, 2006.

[18] A. Ahmed and E. P. Xing, "Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream," *arXiv preprint arXiv:1203.3463*, 2012.

[19] H. Xu, E. Martin, and A. Mahidadia, "Topical establishment leveraging literature evolution," in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 249–252, IEEE Press, 2014.

[20] S. Gupta and C. D. Manning, "Analyzing the dynamics of research by extracting key aspects of scientific papers," in *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp. 1–9, 2011.

[21] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pp. 352–359, Morgan Kaufmann Publishers Inc., 2002.

[22] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 27–34, AUAI Press, 2009.

[23] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 937–946, ACM, 2009.

[24] D. M. Blei and J. D. Lafferty, "Topic models," *Text mining: classification, clustering, and applications*, vol. 10, no. 71, p. 34, 2009.

[25] G. Heinrich, "Infinite LDA implementing the HDP with minimum code complexity," *Technical note, Feb*, vol. 170, 2011.

[26] Y. W. Teh, K. Kurihara, and M. Welling, "Collapsed variational inference for HDP," in *Advances in Neural Information Processing Systems*, pp. 1481–1488, 2007.

[27] M. Levandowsky and D. Winter, "Distance between sets," *Nature*, vol. 234, no. 5323, pp. 34–35, 1971.