Automatic Discovery of Design Task Structure Using Deep Belief Nets

Lijun Lan
Department of Mechanical Engineering,
National University of Singapore,
Singapore 117576, Singapore
e-mail: lijunlan@u.nus.edu

Ying Liu[1]
Mem. ASME
Mechanical and Manufacturing Engineering,
School of Engineering,
Cardiff University,
Cardiff CF24 3AA, UK
e-mail: LiuY81@Cardiff.ac.uk

Wen Feng Lu
Mem. ASME
Department of Mechanical Engineering,
National University of Singapore,
Singapore 117576, Singapore
e-mail: mpelwf@nus.edu.sg

With the arrival of cyber physical world and an extensive support of advanced information technology (IT) infrastructure, nowadays it is possible to obtain the footprints of design activities through emails, design journals, change logs, and different forms of social data. In order to manage a more effective design process, it is essential to learn from the past by utilizing these valuable sources and understand, for example, what design tasks are actually carried out, their interactions, and how they impact each other. In this paper, a computational approach based on the deep belief nets (DBN) is proposed to automatically uncover design tasks and quantify their interactions from design document archives. First, a DBN topic model with real-valued units is developed to learn a set of intrinsic topic features from a simple word-frequency-based input representation. The trained DBN model is then utilized to discover design tasks by unfolding hidden units by sets of strongly connected words, followed by estimating the interactions among tasks on the basis of their co-occurrence frequency in a hidden topic space. Finally, the proposed approach is demonstrated through a real-life case study using a design email archive spanning for more than 2 yr. [DOI: 10.1115/1.4036198]

1 Introduction

The increasing demand for better, faster, and efficient design has driven companies to constantly improve their capabilities of efficient design process management, as a large percentage of defects in product development arise due to an ineffective management of design process [1]. One of the key challenges in design process management is to develop a consistent understanding among design engineers [2]. For this purpose, process modeling has been a prevailing approach that includes various techniques to describe a desired process, define the scope boundary between design tasks, and predict the interaction among design tasks, from flow charts, graphs to agent-based models [3–6]. The current project management practices usually create process models with iterative discussions, which highly rely on the expert evaluation, and many times, the results do not match their initial expectations. The root reasons are usually the uncertainty characteristics of design processes, which is difficult to predict at an early stage, as well as the bias of expert knowledge that is caused by the different backgrounds of designers.

Despite the fact that more appropriate resources are nowadays available due to the increasing application of IT systems in design, study shows that a large percentage of design information required for design process management or process modeling are satisfied only by the designers' individual knowledge base [7]. The design information archived in other forms, such as emails, regular reports, design journals, change logs, and different kinds of social data, are underutilized and completely unexploited in most cases. Actually, all these resources were generated with a specific level of valuable information about the design process. It is essential to make all the audiences contribute to the design process management effectively [8]. However, only a handful of studies [9–11] are found working on extracting useful information from these valuable textual resources for improving the design process understanding.

In order to reduce both the ambiguity introduced by subjective assessment and the time spent in searching for and absorbing information from a large amount of available resources, this paper aims to automatically discover the design task structure from relevant documents collected from completed projects using natural language processing techniques. The ultimate purpose is to help the design engineers to quickly gain insights from the past (one or several completed projects), which would offer great assistance in the management of current and even future projects. The understanding of design tasks, their categorization, and execution patterns harvested from this automatic approach serves as a solid empirical basis for helping the design engineers to make more reasonable and doable decisions.

With the above purpose, this paper mainly focuses on the technical aspect of automatically discovering design tasks and task interactions from design documents. The topic modeling technique which has been widely used in text processing is adopted. In machine learning, a topic model is a type of statistical model that learn latent topics in a collection of documents [12], e.g., latent Dirichlet allocation (LDA) [13], deep belief network (DBN) [14], and Softmax model [15]. For our particular application of design document processing, topic modeling opens the possibility of automated design process analysis for developing a fast understanding of, for example, task executions by transforming raw data in natural language format to task-relevant topics, temporal dynamics of design processes by identifying the changes in task-relevant topics over time, and task interaction patterns by analyzing the co-occurrence frequency of task-relevant topics. In detail, a DBN-based topic modeling approach is proposed in this paper to learn a set of task-relevant topics from design document collections. To deal with documents with different lengths, real-valued units that represent documents in word-frequency vectors are used at the input layer of the proposed DBN topic model. Furthermore, to make the learned topics representative, "label" information that summarizes the central theme of a document, e.g., document title and keywords, is used as the output layer to fine-tune the topic model. Based on the learned topic model, the interaction strength of design tasks are estimated by the co-occurrence frequency of task-relevant topics throughout the document collection.

Section 2 presents a related work about the design process management using process modeling and the art-of-state of topic modeling techniques. Section 3 describes the proposed topic model to discover design tasks, and its utilization to quantify the interaction strength of design tasks. Section 4 reports and discusses the results of applying the proposed approach on a design email archive

[1]Corresponding author.

Contributed by the Design Engineering Division of ASME for publication in the Journal of Computing and Information Science in Engineering. Manuscript received May 14, 2015; final manuscript received March 2, 2017; published online May 16, 2017. Editor: Bahram Ravani.

collected from a real design project. Section 5 discusses some limitations of our current approach. Section 6 gives the conclusions.

2 Related Work

2.1 Modeling Approaches for Design Process Management. In common with all the creative processes, design processes cannot be produced in a strictly linear manner, and the procedure is ever changing throughout the entire lifecycle of the product development. In this context, effective management is of central significance to enable the achievement of product quality, capital in budget, and timeliness. Generally, an effective management to design the process usually commences with ensuring that all designers have a clear understanding of the client's brief, their own scope, and the interactions between design tasks. To address this issue, process modeling has always been an actively researched topic in the filed of design process management. According to the literature, this type of research provides not only a means to represent design process by describing design processes in terms of task dependencies, workflows, inputs, outputs, and so on [8], but also a means to manage the design process by providing a baseline that allows designers to follow a systematic way of thinking about design tasks and their implementations by providing multiple views or single specific view of the entire design system [7].

In the literature, a number of formal design process models have been proposed to provide theoretical support for design process understanding and improvement. Sinha et al. [16] presents a detailed overview of the state-of-the-art in modeling for engineering system design. In this paper, they are classified into three main streams: parameter-based model, task-based model and process-based model.

A typical parameter-based model puts more focus on modeling the connections between bottom-level parameter elements. It is assumed that the parameters of design tasks or functions are a significant aspect of product design, and decisions should be carefully made for these parameters [17]. Generally, these well-founded models confine to a single purpose, which limits their applications in obtaining an overall understanding of a design process. Besides, collecting the detailed information required by this kind of model is a tedious job.

On the basis of parameter-based models, task-based models [18,19] support multiparameter problems via using tasks' parameter connections to quantify design task interaction. The literatures about this kind of model witness the popularity of applying various kinds of design structure matrices (DSMs) on facilitating design process management [20–22]. As a management tool, DSM provides a compact and clear way to present the complex system and highlight the system architecture with multiple views, including process architecture, product architecture, function architecture, and organization architecture. However, the application of DSM in process management is greatly limited by its size. It is difficult to gather reliable information for a DSM with more than ten elements.

Unlike previous two kinds of models, the process-based model aims to provide a good top-level view of the design process with good visibility of different design goals [23]. However, this kind of model is generally with high level of abstraction and cannot support designers with detailed information in process management [7].

Based on the above analysis, we found that although all kinds of process modeling approaches have their potential of being applied to a variety of disciplines, their successful application in design management are often constrained by their capability of integrating reliable and detailed information of design process, which may cause uncounted man-years of meetings, debates, and wastage of both time and money. Besides, it is difficult to model an ever-changing design process only based on experts' evaluations.

In order to address the above problem, this paper aims to propose a way that enables designers to learn design knowledge from the past, which serves as solid and objective experience basis to facilitate process modeling. The arrival of the information-economical world put forward the urgent need of taking advantage of any audience, such as emails and different forms of social data transferred among designers, and regular reports, to support the design process management. Furthermore, the situation that about 20% time of a design is spent by the designers in searching for and absorbing useful information for process improvement speeds this need as well [8]. However, extracting useful information from design documents is an important issue which has not been fully stressed in the literature of design engineering.

2.2 Topic Modeling for Text Document Analysis. In machine learning, a topic model is a type of statistical model for automatically discovering low-dimensional latent topic representation of documents [12]. Intuitively, the per-document word assignments are observed variables, while the topics and per-document topic distributions are hidden variables. Hence, the central problem of topic modeling is using the observed variables to infer the hidden variables.

The most common approach of topic modeling is based on the idea that documents are mixtures of topics, where the word multinomial distributions over a fixed vocabulary correspond to topics. LDA [13] is such a directed graphical model, in which joint distribution is utilized to compute the posterior distribution of the hidden variables. In the literature, LDA is found as the most popular topic model, and its extensions [24–27] also have been applied successfully in finding semantically related words. However, exact inference in these models is difficult, so that the posterior distributions are only computed approximately. Furthermore, these models assume that documents always share the same set of topics.

More recently, neural network-based undirected graphical models are witnessed to outperform LDA models in terms of the low-dimensional latent representation of documents [28–30]. In order to achieve a fast inference, which is difficult for LDA approaches, Gehler et al. [31] used two-layer restricted Boltzmann machines (RBMs) to model word-count vectors as a Poisson distribution. However, they are unable to deal with documents of different length. In order to fix this problem, Hitton and Salakhutdinov [15] proposed the replicated Softmax model. Compared to LDA models, the biggest advantage of these undirected graphical models is that, once trained, it is quite efficient to infer a document's topic feature representation via a simple matrix multiplication.

Even though the inference of the replicated Softmax model [15] is efficient, its representation ability is constrained by the simple network structure of single-hidden layer. In this context, DBN [14] are proposed to learn more complex latent features. A typical DBN consists of one input layer of observations, one output layer of reconstructions of the input data, and several hidden layers. Each hidden layer attempts to reconstruct the input data at a different abstraction level. This deep architecture of DBN allows it to learn more complex topic features than those ones with single network structure. Furthermore, its efficient inference via a simple matrix multiplication makes it possess the capability of outperforming LDA when inferring the topic feature representation of a new document beyond the training dataset.

3 Discover Design Task Structure Via DBN-Based Topic Modeling

Motivated by the great value carried by the available document resources related to design process and the difficulty of mining valuable information from a large set of document resources, this paper proposes a DBN-based topic modeling approach for discovering the design task structure from document collections. Figure 1 illustrates the framework of the proposed approach, where the thin arrow indicates workflows, while the blank arrow indicates input and output flows. In Fig. 1, the starting point is a set of time-

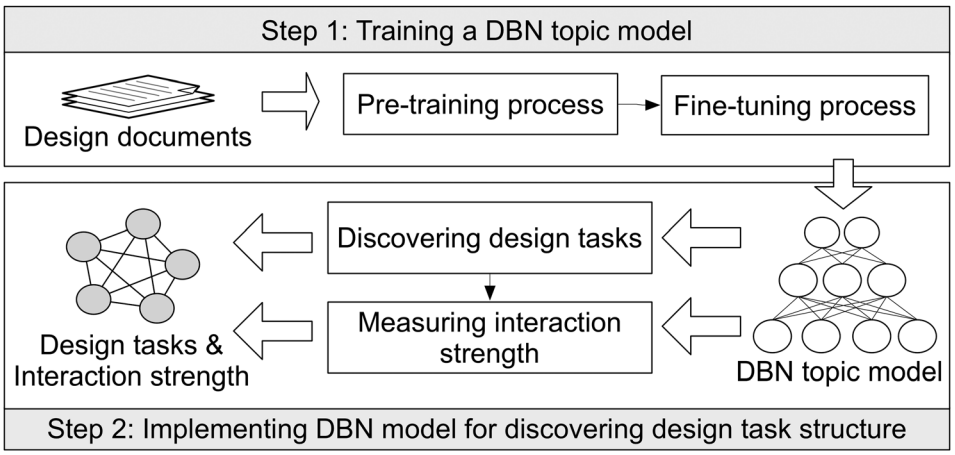

Fig. 1 The framework of discovering design task structure via DBN-based topic modeling

stamped design documents from a real design project. The training process of topic modeling consists of two stages: pretraining and fine-tuning, which create a DBN topic model that learns the latent topics in the input documents. Based on the learned topic model, design tasks are reflected by topics of semantically related words, and task interactions are estimated from the co-occurrence frequency of task-relevant topics. The learned design tasks and their interactions are expected to help designers to gain insight into what design tasks were actually carried out and how they impacted each other in practice, so as to facilitate decision making eventually.

3.1 DBN for Topic Modeling. As shown in Fig. 2, DBN is a probabilistic model composed of one input layer of observations, one output layer of reconstructions of the input data, and several hidden layers [14]. The units of each hidden layer aim to learn the topic representation of the input data (observation) at different abstraction levels. Generally, the abstract topics tend to become more complex as the hidden layer grows.

In order to apply the deep belief networks to extract design tasks from design documents, we consider a document collection D that consists of two parts, i.e., "body" information $\{D_1, \ldots, D_i, \ldots, D_N\}$ and label information $\{Y_1, \ldots, Y_i, \ldots, Y_N\}$. The body information of each document is represented by a word-frequency vector $D_i = (v_1, \ldots, v_j, \ldots, v_M)$, where v_j is the occurrence frequency of the jth word in the body text of document D_i, and M is the vocabulary size of D. The label information is reflected by words in the document title, keywords, and abstract, which summarize a document's central theme and provide significant supplements to text analysis. Taking advantage of such kinds of information to supervise and fine-tune the training process is quite helpful in guaranteeing that the learned latent topics are related to the central theme of a document. Therefore, the label information of a document is defined as $Y_i = (y_1, \ldots, y_j, \ldots, y_M)$, where $y_j \in \{0, 1\}$, and "1" indicates the occurrence of the jth word in D_i 's label information parts. Due to the low frequency of words in the label information, all the words in Y are treated with the same significance. In other words, y_j indicates a word's occurrence in the label information rather than the occurrence frequency in v_j.

By mapping the word-frequency vectors into the visible units in the first layer, the label information into the output units in the highest layer, and the design tasks into the topics captured by hidden units, the problem of discovering design tasks is transformed to find a set of topic features $H = (h_1, \ldots, h_i, \ldots, h_K)$ that not only reconstruct the input data to the largest extent but also mostly connect to the central theme of a document.

As shown in Fig. 2(*a*), the layers of a deep belief network can be split pairwise. Each pair forms a separated restricted RBM, as shown in Fig. 2(*b*), aiming to learn the statistical relationship between the visible units and the hidden units. By this means, the DBN can be greedily trained in a layer-by-layer manner, where the output of the lower-layer RBM is the input data for training a higher-layer RBM. In order to deal with documents with different length and distinguish words with different degrees of contribution, we perform each RBM with a real-valued visible layer and a binary hidden layer.

In detail, a normal distribution is used to model the observed word frequency data V given the hidden topic features H, and a sigmoid function is used to model the hidden topic features H given the observed data V

$$p(v_i|H) = \text{Normal}\left(\frac{\exp\left(\sum_{j=1}^{j=K} w_{ji}h_j + a_i\right)}{\sum_{l=1}^{l=M}\exp\left(\sum_{j=1}^{j=K} w_{jl}h_j + a_l\right)}, 1\right) \tag{1}$$

$$p(h_j = 1|V) = \text{sigm}\left(b_j + \sum_{i=1}^{i=M} w_{ji}v_i\right) \tag{2}$$

where w_{ji} is the symmetric interaction weight between the visible unit (word) v_i and hidden topic h_j, a_i is the bias of the visible unit v_i, and b_j is the bias of hidden unit v_i. The value of visible units stands for the frequency of corresponding words in a document, valued in a range of 0–1. Given a set of topic features, the occurrence frequency over all the words sum up to be one, which is important to deal with documents with different lengths.

The one-step contrastive divergence [32] is adopted to learn the hidden parameters, which are updated by

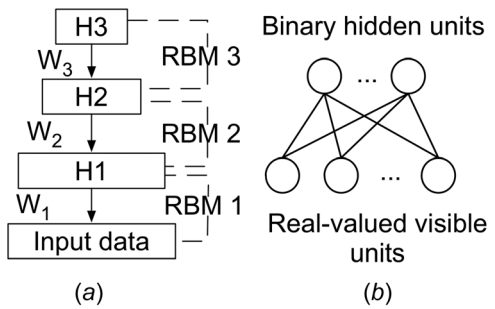

Fig. 2 The architecture of deep belief network (DBN): (*a*) an example DBN with one input layer and three hidden layers, where each pair of succeeded layers is treated as a RBM model and (*b*) the restricted Boltzmann machine (RBM)

$$\triangle w_{ij} = \varepsilon\left(E_{P\text{data}}[v_i h_j] - E_{P\text{recon}}[\hat{v}_i \hat{h}_j]\right) \tag{3}$$

$$\triangle a_i = \varepsilon(E_{P\text{data}}[v_i] - E_{P\text{recon}}[\hat{v}_i]) \tag{4}$$

$$\triangle b_i = \varepsilon\left(E_{P\text{data}}[h_i] - E_{P\text{recon}}[\hat{h}_i]\right) \tag{5}$$

where ε is the learning rate, and $E_{P\text{data}}[v_i h_j]$ is the expectation of the co-occurrence frequency of word v_i and hidden feature h_j given the observed input data, $P\text{data}(v_i, h_j) = p(h_j|v_i)p(v_i)$. Similarly, $E_{P\text{recon}}[\hat{v}_i \hat{h}_j]$ corresponds to the expectation given the reconstructed data via one-step Gibbs sampling.

The training process of DBN consists of two steps: pretraining and fine-tuning. The pretraining step aims to approximate parameters greedily. Each RBM is trained separately. The bottom RBM feeding with the word-frequency vector is expected to learn a set of low-level topic features of a document. The renormalized topic features over the learned posterior distribution $P(H|V)$ are then used as the input data for training a higher-level RBM, which is expected to learn more complex topic features. This layer-by-layer training process is repeated several times to learn a deep belief network in Fig. 3(*a*).

After pretraining, an extra layer of binary units is added to the top of the DBN, as shown in Fig. 3(*b*). The label information $Y_i = (y_1, \ldots, y_j, \ldots, y_M)$ is used to back-propagate the whole network to adjust the weights for learning topic features that are mostly related to document labels. For those documents without the label information, we use the words with high frequency in a document for substitution, considering those with low frequency are less significant to reflect the main idea of a document. The fine-tuning process makes the entire DBN tend to learn the mostly relevant topics in a document.

3.2 Implementing DBN for Discovering Design Task Structure

3.2.1 Discovering Design Tasks. This step advances to use the learned topics to interpret design tasks, which are recorded in design documents. After training, each hidden unit in the topic model is connected to a set of words in the visible layer by weights in W. In turn, the words that are strongly connected reveal the semantic meaning of the corresponding topics, which might refer to a design task in the real word.

In detail, each topic learned by the lowest hidden layer (e.g., H_1 in Fig. 3) is directly represented by words with strongest positive weights to the corresponding hidden unit. Take Fig. 4 as an example, where thick lines indicate strong connections between words and topics. Three words $i \in \{1,\ 2,\ 3\}$ with largest w_{ji} in $w_{j,:}$ are selected to compose Topic_j^1. Similarly, using the low-level topics in place of words in the visible layer, topics learned by higher-level hidden units are represented by groups of strongly connected lower-level topics and tend to convey more complex information about design tasks.

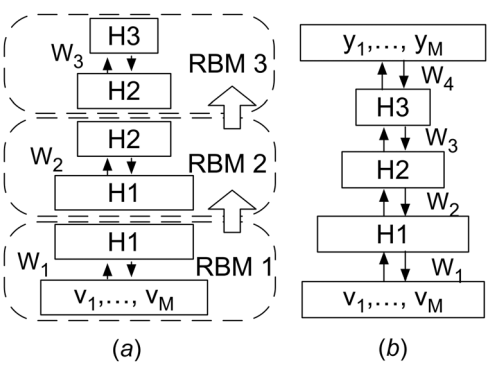

Fig. 3 Training process of DBN topic model: (*a*) pretraining process, in which a stack of RBMs are learned layer by layer and (*b*) fine-tuning process, where an extra layer of label information is added to fine tune the entire network

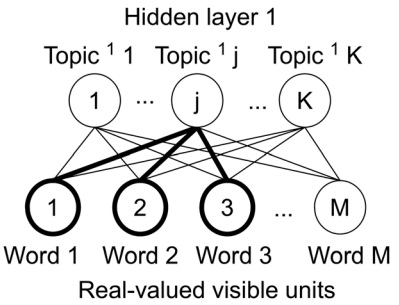

Fig. 4 Illustration of mapping design tasks from hidden topic features. The thick lines indicate words with strongest connections to the *j*th topic.

3.2.2 Measuring the Interaction Strength of Design Tasks. After identifying the relevant design tasks by interpreting each hidden units using the words that strongly connect to it, the trained DBN model is again utilized to assess the interaction strength between design tasks in a hidden topic space. It is natural to consider that design tasks that frequently appear together tend to have stronger connections. Based on this idea, the co-occurrence frequency of design tasks is used as the criterion for measuring their interaction strengths.

For each document, the DBN topic model generates its topic distribution $P(H|V)$ from its word-frequency vector by applying Eq. (2) in a bottom-to-up manner. By mapping latent topics into design tasks, each $P(h_j|V)$ estimates the possibility or frequency that the jth task is recorded in a document. Next, the interaction strength between pairs of design tasks is estimated as

$$IS(h_i|h_j) = \left(\sum_{d=1}^{d=N} P(h_i|V_d)P(h_j|V_d)\right) \Big/ N \tag{6}$$

where N is the size of the document set, and $P(h_i|V_d)P(h_j|V_d)$ computes the co-occurrence frequency of the ith and jth design tasks in the dth document.

4 Case Study

4.1 Experimental Dataset. The case study was conducted on a traffic wave project that aimed to design an Ants transportation (AT) system for tackling the traffic wave problem in the highway system. The project was hosted by a university. The participants primarily consist of six undergraduate students and three professors from three engineering disciplines. Throughout the design process, participants exchanged their opinions and discussed their works via emailing each other in broadcast, one-to-one or one-to-more manners. They were required to always send a copy of any email correspondence to a specific common address. Finally, a set of 569 emails is collected from March 2011 to February 2013 representing a design process from conceptualization to prototyping. All the emails are initially saved in a MS Outlook file and converted to a single XML file.

4.2 Data Preprocessing. The experiments were implemented in Java with the assistance of Apache OpenNLP[1], which is an open source library for processing natural language texts. After manually deleting irrelevant emails, the data were preprocessed by removing meaningless stop-words, performing stemming, and eliminating words that occurred less than two times throughout the

[1]https://opennlp.apache.org/cgi-bin/download.cgi

entire email collection, with the help of Apache OpenNPL. The final vocabulary size is 1968. Words in both email subjects and bodies composed the eventual input data $\{D_1, \ldots, D_i, \ldots, D_N\}$ for pretraining the topic model. Words in email subjects composed the label information $\{Y_1, \ldots, Y_i, \ldots, Y_N\}$ for fine-tuning the topic model.

4.3 Experimental Setup and Evaluation. The performance of the proposed approach is evaluated from two aspects: the effectiveness of full-text document retrieval and the ability for discovering hidden characteristics of the actual design process.

The full-text document retrieval experiment aims to evaluate the influence of the topic model structure on the document retrieval effectiveness. Each email in the training set was used as a query to search those ones with biggest similarity to itself. The content-based similarity of two emails was calculated using the Euclidean distance between their latent topic representations $P(H|V)$. Using the email subject as the evaluation criterion, the document retrieval precision is computed as follows:

$$\text{Precision}(i) = N^{\text{total}}_{\text{correct}}(i)/N_{\text{total}}(i) \tag{7}$$

where N_{total} is the number of emails that have the same subject with the ith email, and $N^{\text{total}}_{\text{correct}}$ is the number of correctly retrieved emails in the top N_{total} ranked relevant emails.

With the help of domain expert knowledge, the ability for design task discovery and task interaction assessment was evaluated by the degree of alignment between automatic findings and participant feedback.

4.4 Results and Discussion

4.4.1 Document Retrieval Evaluation. Figure 5 compares the performances of different DBN models in full-text document retrieval when different numbers of hidden units and hidden layers are selected. Each topic model was trained under the same parameter settings: 2000 iterations for pretraining process, 1000 iterations for fine-tuning process, 0.2 for weight learning rate, and 0.05 for biases learning rate.

By using different numbers of hidden topic units, the average retrieval precision of one-hidden-layer DBN model is shown in Fig. 5(*a*). As seen from the symbol curve in Fig. 5(*a*), the average retrieval precision increases dramatically when the numbers of hidden units are relatively small, but it becomes stable after the number is greater than 50. Based on the well-known experience that more hidden units tend to need more training data and more training time, a moderate number of hidden units is suggested to remain effective in training topic models. For example, the number of hidden units was set to be 50 in the next experiment.

In Fig. 5(*b*), five DBN models with different numbers of hidden layers are compared. As observed from Fig. 5(*b*), compared to the one-hidden-layer model (1630-50), DBNs with two-hidden layers (1630-150-50 and 1630-200-50) improve the precision score from 0.6187 to 0.6438 and 0.6712, respectively. However, different phenomena are found in the two three-hidden-layers DBNs. Accuracies drop to 0.6084 and 0.6147 when a larger number of hidden layers is specified. This conflicting result indicates that the effectiveness of full-text document retrieval is not proportional to the number of hidden topic layers. The insufficient training might be one major reason that the two three-hidden-layers DBNs perform worst. Generally, DBN models of more hidden layers contain more parameters. To guarantee a better result, sufficient training data are required to learn these parameters. Therefore, a moderate number of hidden layers are suggested.

Figure 6 compares one-hidden-layer DBN models with the LDA [20], which is one of the most popular topic models. For fairness, we ran the LDA for 2000 iterations as well, setting the same number of hidden topics as the DBNs. The comparison result in Fig. 6 confirms that DBN outperforms LDA in learning documents' latent topic representation.

4.4.2 Learned Design Tasks. This step aims to inspect that, given a set of design documents, whether the DBN topic model is able to identify meaningful latent topics that uncover design tasks recorded in these documents. Based on the experiment result in Fig. 5, the DBN model of structure 1630-200-50 is selected. For each learned latent topic, also known as the hidden unit in the second hidden layer (the layer of 50 hidden units), we visualized the top 5 words with strongest connections to it and named the corresponding design tasks based on these words. The feedback from project participant reveals that some of the 50 latent topics are truly related to the actual design tasks while some are not.

Due to the space limitation, Table 1 lists six topics that are most relevant to design tasks, which were carried out during this traffic wave project. In the following parts, we will refer design tasks to these topics. For each design task, only words of top-5 strongest connections are listed. According to Table 1, most words

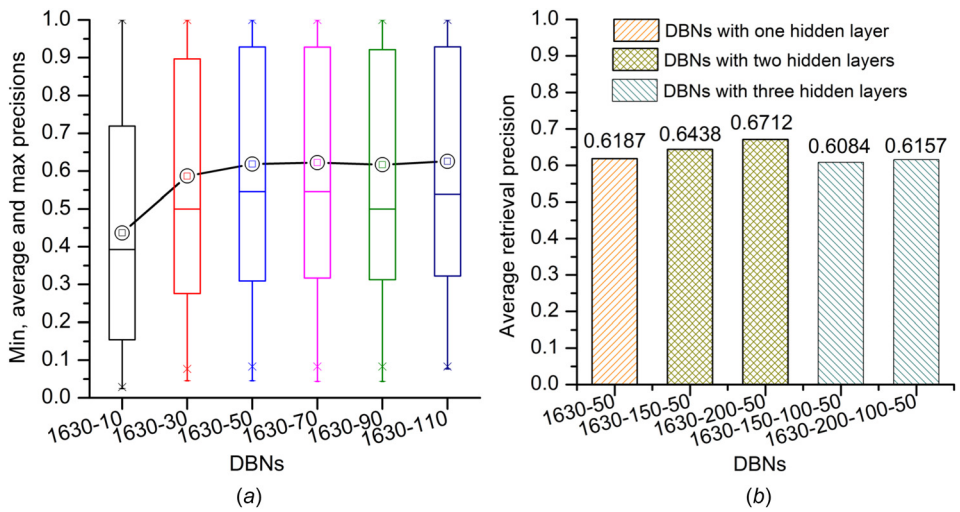

Fig. 5 Document retrieval effectiveness of DBNs: (*a*) comparison of DBNs of one hidden layer but different hidden units and (*b*) comparison of DBNs of the same hidden units in the top layer but different numbers of hidden layers. The DBN structure is indicated in the format of XX-XX, e.g., 1630-50 means a DBN model with one visible layer of size 1630, and one hidden layer of size 50.

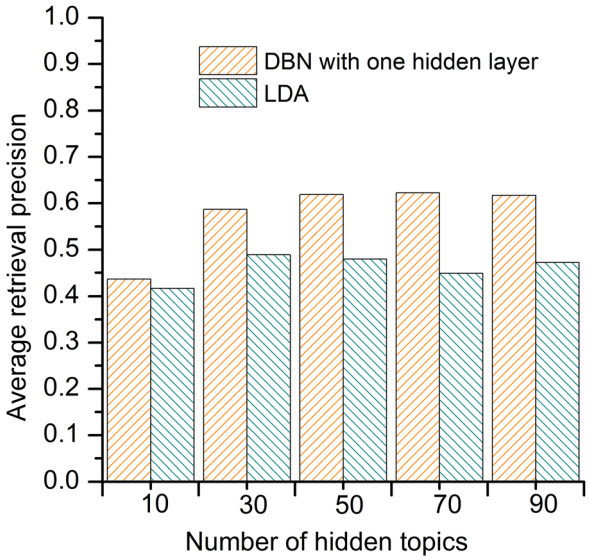

Fig. 6 Document retrieval effectiveness of one-hidden-layer DBNs and LDAs with the same number of hidden topics

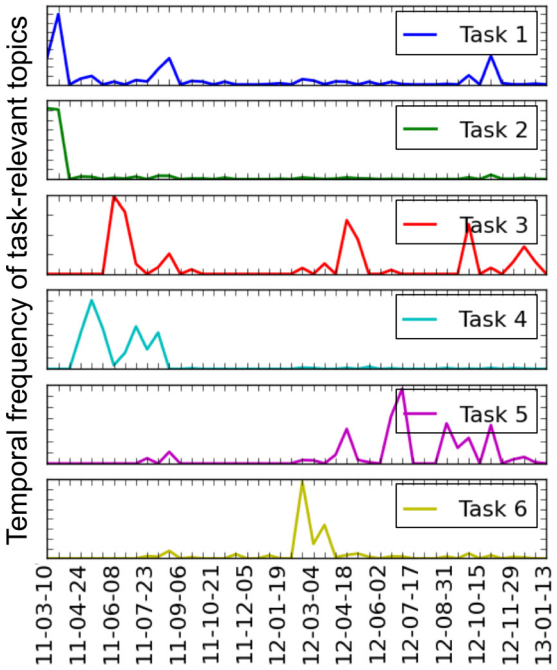

Fig. 7 Temporal frequency of task-relevant topics in Table 1 with a window size of 15 days

associated with each design task are quite intuitive in the sense of conveying a semantic meaning that reflect what were actually done during the design process. Take the six design tasks as an example, namely XXX project proposal, concept paper submission, ASME conference paper, IRB application, traffic data collection, and simulation software. According to the feedback from one core participant, the traffic wave project is only a part of the XXX project, which consists of several subprojects. At the beginning, each subproject was required to submit a project proposal. Next, a detailed concept paper about their ideas and plans was completed with the efforts of all the participants of the traffic wave project after several modification iterations, which is reflected by task 2. In the middle stage, an unexpected task was conduced to obtain some supporting documents from a significantly relevant department, which took quite a long time to finish the IRB application. After developing the core techniques that are not shown in Table 1, real-life traffic data were fetched from the traffic department and utilized to evaluate the developed Ants transportation system on several simulation platforms, one of which is named as Paramics. Finally, this project was ended with writing and publishing an ASME paper that summarized the main work and achievement of this project.

In order to track the regions of the timeline when students were truly working on the different tasks, Fig. 7 plots the temporal frequency of the six task-relevant topics in Table 1 with a window size of 15 days. Again, the timeline of each task in Table 1 aligns well with the above feedback. It can be seen that students first conducted on the project proposal issue (task 1) and achieved a concept paper (task 2) during the first month after the project started out. By the second month, students proceeded to obtain the IRB support (task 4) before they could advance to the technical part, which took them about 4 months. The traffic data collection (task 5) and simulation software purchase (task 6) were started out almost simultaneously after about 10 months. However, students spent much longer time in finding out what types of data they could get from the traffic department and processing these data.

Above observations demonstrate the ability for uncovering design tasks and their temporal dynamics from collected documents. The findings were evaluated by the project participant. Their feedback

Table 1 Illustration of selected design tasks learned by DBN topic model. Each topic is represented by five words with strongest connection to it. The probability column displays the weights connecting words and topics. For privacy reasons, XXX is used in place of the names of organizations and persons.

Words	Probability	Words	Probability	Words	Probability
Task 1 (XXX project proposal)		Task 2 (concept paper submission)		Task 3 (ASME conference paper)	
XXX	0.591	Concept	0.598	Revise	0.706
Meeting	0.291	Submission	0.556	ASME	0.315
Proposal	0.245	Revise	0.276	Dates	0.268
Project	0.230	Paper	0.267	Congress	0.259
Importance	0.022	Conference	0.223	Ants	0.190
The words	Probability	Words	Probability	Words	Probability
Task 4 (IRB application)		Task 5 (traffic data collection)		Task 6 (simulation software)	
Application	2.496	Traffic	0.514	PARAMICS	0.527
IRB	2.696	AYE	0.499	Simulation	0.293
Review	2.235	Data	0.492	Key	0.256
XXX	0.597	Project	0.482	Software	0.193
Form	0.566	Program	0.433	Wei	0.137

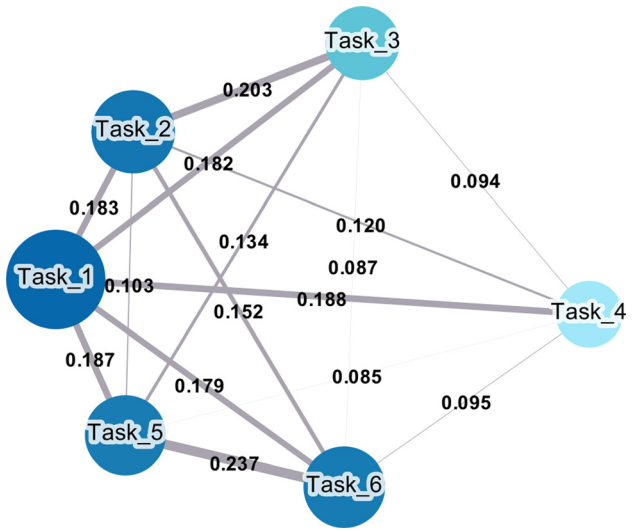

Fig. 8 Illustration of interaction strengths between selected design tasks

revealed strong positive comments to the results, which uncovered the actual design process from concept generation, methodology and key techniques development, experiment data collection, to experiment validation via simulation.

4.4.3 Learned Design Task Interactions. After identifying design tasks, this step advances to answer the question that how design tasks had interacted with each other in practice. For each email, we generated its design task distribution by feeding the trained DBN model with its word-frequency vector, then computed design tasks' co-occurrence frequency based on Eq. (6).

Figure 8 illustrates the interaction strength between above six tasks, where nodes indicate tasks, the size of nodes reflect the overall interaction between one task and all others, and the thickness of edges are related to the strength connecting tasks. From Fig. 8, one notable observation is that task 1 (XXX project proposal in Table 1) might have interacted with all others strongly and equally. This finding is not difficult to explain. Because all the initial design ideas were generated in this task, it is natural that all the remaining tasks had connections with it more or less. Edges connecting task 4 (IRB application) show an exactly inverse interaction pattern, namely, task 4 only has strong connections to two tasks, tasks 1 and 2, with a strength of 0.188 and 0.120, respectively. This observation is consistent with the feedback that task 4 is not a part of the design project itself, but required to get support from relevant departments based on the results of tasks 1 and 2. The strongest interaction, valued at 0.237, is found between tasks 5 and 6. This is validated by the relevant emails that the two tasks were carried out concurrently, and both were related to validating the finally developed Ants transportation system. Taken together, the above findings align well with the feedback of the project participant, which proves the ability of our approach for task structure discovery and analysis.

5 Discussion

The purpose of this study is to extract useful design information and knowledge from textual data to help in developing an understanding of historical operations. Two significant aspects have been considered: design tasks which are mapped from frequent topics in design documents, and task interaction strengths which are estimated according to the co-occurrence frequency of corresponding topics. The experiment results demonstrate a good alignment between the automatic design information and expert evaluation. This proves that our approach can not only provide assistance for design engineers in obtaining deep insight into the behavior of actual design processes but also save both time and labor required by the traditional human analysis.

Despite the above insightful contributions, some limitations still remain. First, even though the learned topics could reveal some design tasks in real world, the words composing a topic are difficult for interpretation, especially for novices. Take task 1 in Table 1 as an example. Based on words, i.e., "meeting," "proposal," "project," and "importance," participants of this project can easily recollect corresponding tasks, but it might be difficult for novices to connect them to a real-world task. This is caused by the learning mechanism of topic models, which discover abstract "topics" only based on the statistics of words, overlooking their occurrence order. Second, although the co-occurrence frequency of topics can reveal the task interaction to some extent, it is not sufficient to explain how design tasks interacted. One most significant reason is that the complex interaction between tasks is jointly determined by multiple process variables. Consequently, identifying these process-related variables in design document is critical for estimating tasks' interaction strength more comprehensively and correctly. Both limitations drive us to extract and analyze design information with a more fine-grained granularity in our future work.

6 Conclusions

Digital design documents provide potentially useful sources of valuable experience that would assist decision-making in future projects. In this paper, we proposed a DBN topic modeling approach to discover design tasks and estimate their interaction strengths from textual design data. The case study was conducted on a set of emails collected from a real-life design project. Further, the experimental results show that our approach produces identical results with the feedback from project participants, which proves the metrics of our approach in helping engineers to get deep insight into a historical project. Discussions on relevant concerns also highlight some future research possibilities, e.g., fine-grained task discovery and comprehensive interaction assessment, using more advanced learning techniques in text mining, natural language processing, machine learning, and statistics.

Nomenclature

a_i = bias of visible unit v_i
b_j = bias of hidden unit h_j
D_i = word-frequency representation of the ith document
H = vector of hidden units
h_i = ith hidden unit
M = vocabulary size
N = number of design document
N_{total} = number of emails that have the same subject with the ith email
$N_{\text{correct}}^{\text{total}}$ = number of correctly retrieved emails in the top N_{total} ranked relevant emails
$P(h_j/V_d)$ = possibility that the jth design task is recorded in the dth document
v_i = occurrence frequency of the ith word in a document
w_{ji} = interaction weight between visible unit (word) v_i and hidden unit h_j
y_i = "1" indicates the occurrence of the jth word in a document's label information
Y_i = word-occurrence representation of the ith document's label information

References

[1] Formoso, C. T., Tzotzopoulos, P., Jobim, M. S., and Liedtke, R., 1998, "Developing a Protocol for Managing the Design Process in the Building Industry," 6th Annual Conference of the International Group for Lean Construction (IGLC-6), Guaruja, Brazil, Aug. 13–15.
[2] Qamar, A., Paredis, C. J. J., Wikander, J., and During, C., 2012, "Dependency Modeling and Model Management in Mechatronic Design," ASME J. Comput. Inf. Sci. Eng., **12**(4), p. 041009.
[3] Mou, G., and Tanik, M., 2002, "Transdisciplinary Project Management Through Process Modeling," J. Integr. Des. Process Sci., **6**(3), pp. 45–62.

[4] Chin, K. S., Mok, C. K., and Zu, X., 2007, "Modeling and Performance Simulation of Mould-Design Process," Int. J. Adv. Manuf. Technol., **34**(3), pp. 236–251.
[5] Chen, C.-H., Ling, S. F., and Chen, W., 2003, "Project Scheduling for Collaborative Product Development Using DSM," Int. J. Project Manage., **21**(4), pp. 291–299.
[6] Karniel, A., and Reich, Y., 2009, "From DSM-Based Planning to Design Process Simulation: A Review of Process Scheme Logic Verification Issues," IEEE Trans. Eng. Manage., **56**(4), pp. 636–649.
[7] Baxter, D., Gao, J., Case, K., Harding, J., Young, B., Cochrane, S., and Dani, S., 2007, "An Engineering Design Knowledge Reuse Methodology Using Process Modelling," Res. Eng. Des., **18**(1), pp. 37–48.
[8] De Mera Sánchez, P. D., Gaya, C. G., and M. Á. S. Peréz, 2013, "Standardized Models for Project Management Processes to Product Design," Proc. Eng., **63**, pp. 193–199.
[9] Cheong, H., and Shu, L. H., 2012, "Automatic Extraction of Causally Related Functions From Natural-Language Text for Biomimetic Design," ASME Paper No. DETC2012-70732.
[10] Liu, Y., Liang, Y., Kwong, C. K., and Lee, W. B., 2010, "A New Design Rationale Representation Model for Rationale Mining," ASME J. Comput. Inf. Sci. Eng., **10**(3), p. 031009.
[11] Mathieson, J., Miller, M., and Summers, J., 2011, "A Protocol for Connective Complexity Tracking in the Engineering Design Process," DS 68-7: 18th International Conference on Engineering Design (ICED 11), Impacting Society Through Engineering Design, Vol. 7, Human Behaviour in Design, Lyngby/Copenhagen, Denmark, Aug. 15–19, pp. 492–500.
[12] Steyvers, M., and Griffiths, T., 2007, "Probabilistic Topic Models," Handb. Latent Semantic Anal., **427**(3), pp. 427–448.
[13] Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003, "Latent Dirichlet Allocation," J. Mach. Learn. Res., **3**(5), pp. 993–1022.
[14] Bengio, Y., 2009, "Learning Deep Architectures for AI," Found. Trends Mach. Learn., **2**(1), pp. 1–27.
[15] Hinton, G. E., and Salakhutdinov, R., 2009, "Replicated Softmax: An Undirected Topic Model," 23rd Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, Dec. 7–10, pp. 1607–1614.
[16] Sinha, R., Paredis, C. J. J., Liang, V.-C., and Khosla, P. K., 2001, "Modeling and Simulation Methods for Design of Engineering Systems," ASME J. Comput. Inf. Sci. Eng., **1**(1), pp. 84–91.
[17] Browning, T. R., 2001, "Applying the Design Structure Matrix to System Decomposition and Integration Problems: A Review and New Directions," IEEE Trans. Eng. Manage., **48**(3), pp. 292–306.
[18] Steward, D. V., 1981, "The Design Structure System: A Method for Managing the Design of Complex Systems," IEEE Trans. Eng. Manage., **EM-28**(3), pp. 71–74.
[19] Wynn, D. C., Eckert, C. M., and Clarkson, P. J., 2006, "Applied Signposting: A Modeling Framework to Support Design Process Improvement," ASME Paper No. DETC2006-99402.
[20] Kumar, P., and Mocko, G., 2007, "Modeling and Analysis of an Ontology of Engineering Design Activities Using the Design Structure Matrix," ASME Paper No. DETC2007-35634.
[21] Shi, Q., and Blomquist, T., 2012, "A New Approach for Project Scheduling Using Fuzzy Dependency Structure Matrix," Int. J. Project Manage., **30**(4), pp. 503–510.
[22] Othman, M., Bhuiyan, N., and Kong, L., 2011, "Developing a Dynamic Wheelchair Using the Design Structure Matrix Method," Concurrent Eng., **19**(3), pp. 235–243.
[23] Cheng, F., Li, H., Wang, Y.-W., Skitmore, M., and Forsythe, P., 2013, "Modeling Resource Management in the Building Design Process by Information Constraint Petri Nets," Autom. Constr., **29**, pp. 92–99.
[24] Blei, D. M., and Lafferty, J. D., 2006, "Dynamic Topic Models," 23rd International Conference on Machine Learning (ICML), Pittsburg, PA, June 25–29, pp. 113–120.
[25] Yanning, Z., and Wei, W., 2014, "A Jointly Distributed Semi-Supervised Topic Model," Neurocomputing, **134**, pp. 38–45.
[26] Boyd-Graber, J. L., and Blei, D. M., 2009, "Syntactic Topic Models," 22nd Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, Dec. 8–10, pp. 185–192.
[27] Weinshall, D., Levi, G., and Hanukaev, D., 2013, "LDA Topic Model With Soft Assignment of Descriptors to Words," 30th International Conference on Machine Learning (ICML), Atlanta, GA, June 16–21, pp. 1748–1756.
[28] Larochelle, H., and Lauly, S., 2012, "A Neural Autoregressive Topic Model," 26th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, Dec. 3–8, pp. 2708–2716.
[29] Wan, L., Zhu, L., and Fergus, R., 2012, "A Hybrid Neural Network-Latent Topic Model," 15th International Conference on Artificial Intelligence and Statistics (AISTATS), La Palma, Spain, Apr. 21–23, pp. 1287–1294.
[30] Abdelbary, H. A., Elkorany, A. M., and Bahgat, R., 2014, "Utilizing Deep Learning for Content-Based Community Detection," Science and Information Conference (SAI), London, Aug. 27–29, pp. 777–784.
[31] Gehler, P. V., Holub, A. D., and Welling, M., 2006, "The Rate Adapting Poisson Model for Information Retrieval and Object Recognition," 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, June 25–29, pp. 337–344.
[32] Hinton, G. E., Osindero, S., and Teh, Y.-W., 2006, "A Fast Learning Algorithm for Deep Belief Nets," Neural Comput., **18**(7), pp. 1527–1554.