Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 54-1 (2021) 1200–1205

Automatic Identification of Bottleneck Tasks for Business Process Management using Fusion-based Text Clustering

Junya Tang*, Li Li*
Ying Liu and Kuo-Yi Lin***

* School of Electronics and Information Engineering, Tongji University, 201804, Shanghai, China (Tel: 13817921579;
e-mail:1610453@ tongji.edu.cn; lili@tongji.edu.cn; 19603@tongji.edu.cn).
** School of Engineering, Cardiff University, Cardiff, United Kindom (e-mail: LiuY81@cardiff.ac.uk)

Abstract: With the arrival of the industrial big data era, it offers unprecedented opportunities for machine learning to intelligently uncover hidden tasks and restore the entire underlying process for business process modelling. While recent studies, e.g., process mining and ontologies, have advanced the research agenda of business process modelling and management, identifying a bottleneck task automatically needs more in-depth research. In this paper, a text mining-based bottleneck task identification approach is proposed. Firstly, to extract tasks from documents in different lengths, a dynamic sliding window is introduced to the biterm topic model. The sliding window size is adjusted according to document length during biterm selection process to ensure the two words in biterm comes from a context. Secondly, a fusion-based clustering algorithm is studied to uncover business tasks. The improved biterm topic model and the Doc2vec model are used to train two document vectors and then calculate two distances. The linear fusion of these two distances is used as the metric of clustering. Thirdly, the temporal frequency of each task at different periods is calculated to show the timeline and abnormal occurrence of tasks to identify bottleneck tasks. The proposed approach is evaluated using a data set containing the execution of a multi-year multidisciplinary student design project. The experiment results show the approach can effectively identify bottleneck tasks without manual intervention.

Keywords: Business process modelling, Workflow mining, Topic model, Text clustering

1. INTRODUCTION

The diversity and customization of products and services put forward higher requirements for manufacturers and service providers. Business process management and modelling offer organizations unprecedented support to meet customers via reusing embedded knowledge from previous projects, such as design experience, control strategy, and product parameters (Janssenswillen G et al., 2019). Knowledge extraction and visualization is the foundation of advanced business process management and modelling. Significant studies have been conducted in process mining and ontologies for knowledge extraction, such as task identification, workflow reconstruction from structured business process event logs (Bozorgi Z D et al., 2020). However, enough attention has not been given to the automatic identification of bottleneck tasks from unstructured data. Bottleneck tasks are tasks taking a long time or causing rework of other tasks, which is the key to the smooth execution of a process. Most research on bottleneck tasks focuses on the discovery of short-loop structures via Petri nets.

Besides, most work conducted in the above areas can only be applied to structured data such as event logs. Other channels supporting the execution of the business process also collect amounts of valuable data. Several studies recently extended the scope to amounts of unstructured data sources, such as emails, meeting minutes, and conversation records (Elleuch M et al., 2020b). However, most of these studies also rely on a manual intervention, which is easy to cause different understanding among different engineers (Lijun Lan et al., 2017).

This paper mainly focuses on the automatic identification of bottleneck tasks from unstructured business process data and proposed a text-clustering based bottleneck task identification approach. Focusing on documents in different lengths, a dynamic sliding window was introduced to biterm topic model (BTM). To enhance the features extracted from the document, combining the statistical model and the neural network model to calculate the fusion distance as the distance metric of document clustering. Based on the learned topic models and divided document clusters, the lasting time of each task is estimated according to the occurrence frequency, and rework tasks can be found by the reconstructed workflow. According to these two indicators to identify bottleneck tasks.

The subsequent sections are structured as follows. In Section 2, a literature review is presented. The approach is proposed in Section 3. A case study that automatically restores task flow and identifies bottleneck tasks is revealed in Section 4. Finally, the conclusion is given in Section 5.

2. LITERATURE REVIEW

10.1016/j.ifacol.2021.08.142

2.1 Business Process Modelling Approaches

Business process management provides a possibility to bridge the gap between data and process science. Process mining methodologies such as process discovery, process enhancement have promoted business management by discovering models from event log data (Y Liu et al., 2020; Lijun Lan et al., 2018). Traditionally, process mining always focuses on workflow discovery, that is, discovering tasks and the execution patterns between them from structured event logs (Y Liu et al., 2020). Traditional process mining provides strong support for business management; however, given the unstructured process data such as emails, meeting minutes, and conversation records, traditional process mining methodologies could not be applied directly (Elleuch M et al., 2020a). Recently, there have been several researchers tending to unstructured data. Lijun Lan et al. (2018) proposed a deep belief net (DBN) to automatically extract the design task structure from email data collected during a design project and built a hierarchical process model. This model displayed the design process behind amounts of design documents from different perspectives and provided a consistent understanding to different designers. Elleuch M et al. (2020a) introduced a pattern discovery-based approach to discover frequent activities from email logs, which reduced human intervention. However, these models are still abstract for staff to quickly grasp the limitation of a business process.

Some researchers paid attention to the bottleneck in the business process, including bottleneck prediction and bottleneck identification. Bottlenecks always exit in mixed multiple-concurrency short-loop structures (Sun H W et al., 2020). Petri nets (Zhu F et al., 2019) and unbounded Petri nets (Lu F et al., 2019) are traditional methodologies to identify it. Based on Petri nets, local task search is also used to enhance bottleneck discovery via local event logs (Vázquez-Barreiros B et al., 2016). Recently, deep neural networks (DNN) (Huang B et al., 2019) are also used to predict bottlenecks based on real-time data collected by the internet of things (IoT). Henri Boessenkool (Boessenkool H et al., 2018) proposed a three-phased task analysis approach that not only identifies but also quantifies bottlenecks in the telemanipulated maintenance process. Some other approaches, such as complex network analysis, series analysis, are also developed.

2.2 Topic Discovery for Document Analysis

The topic model is a common strategy for discovering the latent semantics and giving the topic representation of documents. The most common topic modelling approach is latent dirichlet allocation (LDA) (Blei et al., 2003), which got a good performance in long texts. LDA has several extensions, such as light LDA. However, all these models are based on a hypothesis that documents are mixtures of topics and documents always share the same set of topics. This hypothesis is not always reasonable, especially for short texts.

BTM (Yan X et al., 2013) takes the words biterm set as modelling objects, and the topic of documents depends on the biterms it contains, which can solve the above problem. Two-layer restricted Boltzman machines (Gehler et al., 2006), replicated Softmax model (Hinton et al., 2009), and DBN (Bengio and Y., 2009) are also proposed for the above problem via modelling word-count vectors. However, these models focus on documents of the same length. Lijun Lan et al. (2018) used real-value units to represent documents in different lengths in word-frequency vectors at the input layer of the DBN topic model. Gupta P et al. (2019) proposed neural autoregressive topic models combined with external knowledge. However, appropriate external knowledge does not always exist.

2.3 Document Clustering

Document clustering is always used to discover hot topics of short news (Cheng X et al., 2014). It is always based on the features extracted from texts. Zheng Y et al. (2015) proposed a TF-IDF&K-means algorithm in which the term frequency-inverse document frequency (TF-IDF) was used to represent features, and then K-means was used to catch news topics. However, TF-IDF only focuses on the frequency and ignores the textual semantic information. Li W et al. (2016) developed a hot topic detection algorithm based on BTM and K-means to enhance the semantic information of vectored texts. Yamin W and Yue H (2016) proposed an algorithm based on BTM and TF-IDF to combine frequency and semantic information of words. The vector-matrix represented by TF-IDF is local sparse for documents in different lengths. Compared to TF-IDF, Doc2Vec can reduce the dimensionality of texts in different lengths and directly calculate the similarity of the texts. In this paper, the advantages of the biterm topic model with a dynamic sliding window (BTMDW) distinguishing meaning and Doc2Vec dealing with texts in different lengths and polysemy are combined in the distance metric of K-means.

3. METHODOLOGY

Motivated by the great value of available business process document resources and the difficulty of extracting valuable knowledge from these document resources, this paper proposed a text clustering-based bottleneck tasks automatic identification approach. The framework of the proposed approach was illustrated in Fig.1.

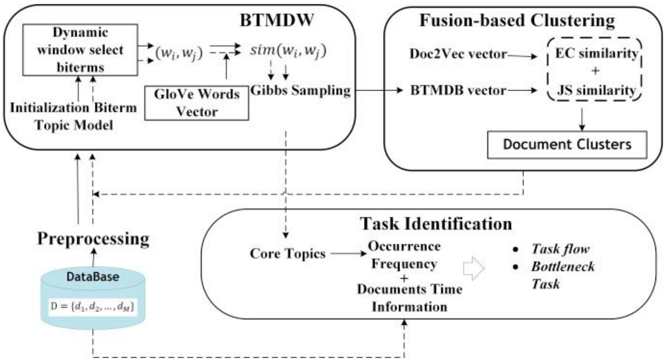

Fig. 1. The framework of business bottleneck task automatic identification.

The framework includes three parts, an improved biterm topic model with a dynamic sliding window (BTMDW) extracts topics from documents and represents documents via vectors. The fusion-based clustering model divides input data into several clusters according to fusion distance. Task identification reconstructs design task flow according to occurrence time of topics and identifies bottleneck task.

3.1 Biterm Topic Model with Dynamic Sliding Window

BTM was proposed for short texts (Yan X et al., 2013), and it solved the word sparsity problem via using biterms instead of words. Biterm denotes an unordered word pair co-occurring in a short context (Cheng X et al., 2014). Here a short context refers to a small part over a term sequence (Cheng X et al., 2014). Simply, in short texts with limited document length, each document is taken as an individual context unit. However, with the text length increasing, the hypothesis is invalid, and the number of biterms will overgrow. The biterm formed between two far apart words is almost worthless and will cause redundancy in calculations. Due to the length of business process documents is uncertain, BTM can not be applied directly to business process documents. An improved BTM with a dynamic sliding window (BTMDW) was proposed to tackle this problem.

Firstly, to select biterms from documents in different lengths, a dynamic sliding window is introduced to BTM. The size of the sliding window is adjusted according to the length of the document. Biterm selection process with a sliding window is shown in Fig.2. After the sliding window, some biterm candidates are extracted. However, some biterms are worthless due to the low correlation between the word pair. Then, Glove is used to train word vectors and calculate the cosine similarity between the word pair in biterm candidates.

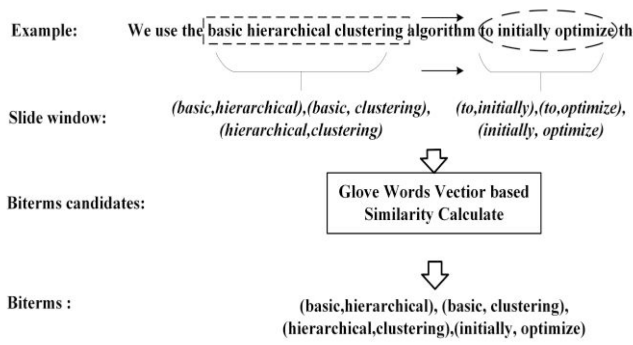
Example:
We use the basic hierarchical clustering algorithm to initially optimize the...
Slide window:
(basic,hierarchical),(basic, clustering),
(hierarchical,clustering)
(to,initially),(to,optimize),
(initially, optimize)
Biterms candidates:
Glove Words Vectior based Similarity Calculate
Biterms :
(basic,hierarchical), (basic, clustering),
(hierarchical,clustering),(initially, optimize)

Fig. 2. Biterm selection (slide window size=1)

After biterm selection, Gibbs sampling was used to estimate the multinomial distribution parameters ϕ and θ. For biterm b_i , sampling its topic z_i according to the following conditional distribution (Cheng X et al., 2014):

$$P(z_i = k \mid z_{-i}, B) \propto (n_{-i,k} + \partial) \cdot \frac{(n_{-i,w_{i,1}|k} + \beta)(n_{-i,w_{i,2}|k} + \beta)}{(n_{-i,\cdot|k} + W\beta + 1)(n_{-i,\cdot|k} + W\beta)} \tag{1}$$

z_{-i} denotes the topic distribution for all biterms except b_i. $n_{-i,k}$ is the number of biterms assigned to topic k except b_i. $n_{-i,w_{i,1}|k}$ denotes the number of times word w assigned to topic k except b_i , and $n_{-i,\cdot|k} = \sum_{w=1}^{W} n_{-i,w|k}$. After enough iterations, ϕ and θ are estimated as follows:

$$\phi_{k,w} = \frac{n_{w|k} + \beta}{n_{\cdot|k} + W\beta} \tag{2}$$

$$\theta_k = \frac{n_k + \partial}{N_B + K\partial} \tag{3}$$

n_k denotes the number of biterms in each topic k . $n_{w|k}$ denotes the number of times that each word w is assigned to the topic k.

3.2 Fusion Distance-based Text Clustering

For the K-means clustering algorithm, selecting a distance metric is essential, so the BTMDW-DOC-K-means text clustering algorithm is proposed and applied to the business process. First, business process documents are collected and preprocessed. Then, they are modelled via BTMDW and Doc2Vec. After BTMDW modelling, business process documents can be represented as vectors and JS divergence is adopted to calculate the similarity between documents. At the same time, Euclidean distance is adopted to calculate the similarity between documents represented via Doc2Vec modelling. Finally, fuse distance based on statistical model BTMDW and distance based on neural network model Doc2Vec and apply fusion distance to K-means.

The document d_i can be represented as a vector via BTMDW modelling (Cheng X et al., 2014):

$$d_{i\_BTMDW} = \{p(z_1 \mid d_i), p(z_2 \mid d_i), ..., p(z_k \mid d_i)\} \tag{4}$$

The distance between two documents d_i and d_j is calculated by JS divergence between the two document vectors d_{i_BTMDW} and d_{j_BTMDW} . The distance based on BTMDW modelling and JS divergence is as

$$DIS_{JS}(d_{i\_BTMDW}, d_{j\_BTMDW}) = \frac{DIS_{KL}(d_i \parallel \frac{d_i + d_j}{2}) + DIS_{KL}(d_j \parallel \frac{d_i + d_j}{2})}{2} \tag{5}$$

Doc2Vec (Park S et al., 2019) model maps documents into the vector space to maintain the semantic similarities using the context of a word given. M denotes the vector size. Then, the document d_i can be represented by a M dimension vector via Doc2Vec modelling,

$$d_{i\_Doc2Vec} = \{v_1(d_i), v_2(d_i), ..., v_M(d_i)\} \tag{6}$$

The similarity between the two document vectors $d_{i\_Doc2Vec}$ and $d_{j\_Doc2Vec}$ is calculated by Euclidean distance. The distance based on Doc2Vec modelling and Euclidean distance is as:

$$DIS_E(d_{i\_Doc2Vec}, d_{j\_Doc2Vec}) = \| d_{i\_Doc2Vec} - d_{j\_Dov2Vec} \|_2 \quad (7)$$

After training the documents vector, K-means is used to cluster input documents. Firstly, selecting K texts from document collection D as the initial centres of K clusters. Then, assigning each text to the most similar cluster via calculating fusion distance between text and centres. After that, updating the cluster centres and repeat the above process. The fusion distance metric between two documents d_i and d_j is a linear combination of DIS_{JS} and DIS_E, and the combination coefficient is λ,

$$DIS(d_i, d_j) = \lambda \cdot DIS_{JS} + (1-\lambda) DIS_E \quad (8)$$

3.3 Bottleneck Task Identification

Task and task execution time need to be got first to identify bottleneck task. After fusion distance-based text clustering, several clusters of documents can be obtained. Task extraction assumes that each document cluster contains a core topic considered a task.

Firstly, the core topic of each document cluster is extracted via the first part BTMDW. Topic word distribution and topic distribution can be estimated via the function (2) and function (3). According to the topic word distribution trained by BTMDW, take the top four words to form the topic. Secondly, the lasting time participants are working on each task is tracked to restore the task flow and identify bottleneck tasks. The occurrence frequency of task-relevant topics within a set period, such as one week, is calculated to show the timeline of each task. Besides time-consuming, rework is another indicator to identify bottleneck tasks. From the timeline, the question of how each task causes other tasks to rework can be answered.

4. CASE STUDY

4.1 Experimental Dataset

The case study was conducted on a real design project hosted by a university. The project aimed to track the traffic wave problem in the highway system via designing an Ants transportation system. During the design process, all participants, including several students and three professors, notified project activities, discussed works, exchanged ideas via email. Copies of all emails were required to be sent to a common address. During the two years of the entire project, 569 emails were collected and stored as an XML file, including activities, resources and personnel interactions during the design process.

4.2 Data Pre-processing

The experiments were implemented in python. After extracting the emails from the XML file and manually filtering out incorrect emails, the documents were preprocessed via the NLTK library in python. After removing meaningless stop-words, performing stemming, the final vocabulary size is 2928.

4.3 Bottleneck Task Identification

This experiment aims to extract the design process tasks and then find the bottleneck task during this project. We compare results with the DBN topic model proposed by Li Jun et al. (2018) on the same data set. Two metrics are used to evaluate the performance. One is the similarity between the extracted task flow and the planned task flow, and the other is bottleneck tasks identification.

The first step is setting two parameters of BTMDW-DOC-K-means, cluster number K and the fusion coefficient λ. According to prior knowledge, the project usually has five to nine tasks, so K was selected from 5 to 9. And fusion coefficient λ was set from 0.2 to 0.9. Three measures silhouette-score (S) score, Calinski-Harabaz (CH) score, and Davies-Bouldin (DBI) score are used for parameter selection. The parameters maximize S score and CH score and minimize DBI score are selected. The result is shown in Fig.3, and we can see that when K is 7 and λ is 0.7, the S score and CH score are maximum, and DBI score is minimum.

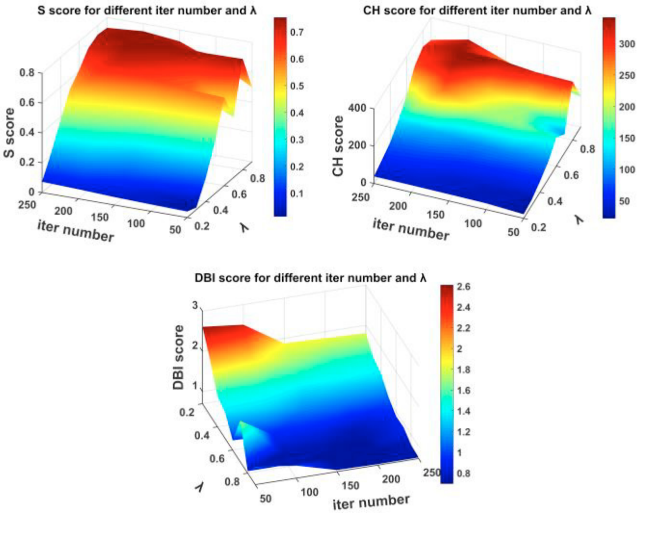

Fig. 3. S value, CH value and DBI value for different cluster numbers K and λ value.

Then, seven document clusters are obtained by clustering the input documents, and the core topics of these seven document clusters are extracted via BTMDW. Each topic is represented by its top four words, as shown in Table 1. The extracted tasks are understood as concept paper and student group, project proposal, transportation system design, software application and system simulation, research paper submission and presentation, certain vehicle and video presentation, and entire program optimization.

DBN topic model-based approach (Lijun Lan et al., 2017) extracted 50 latent topics from given documents. Then six topics that are most relevant to design tasks are selected according to experience. Each topic is represented by its top five words. The extracted tasks are understood as project proposal, concept paper submission, conference paper, IRB application, traffic data collection and simulation software.

Table 1. Top Words of Tasks Selected by BTMDW-DOC-K-means

Top Words of Tasks Selected by BTMDW-DOC-K-means			
Task 1	*Task 2*	*Task 3*	*Task 4*
concept	traffic	system	software
paper	vehicle	design	simulation
student	project	transportation	system
group	proposal	draft	application
Task 5	*Task 6*	*Task 7*	
research	vehicle	program	
paper	certain	entire	
presentation	video	optimize	
submission	presentation	submission	

According to the generation time of the corresponding document of the task, the task flow of this project can be restored. We compared the task flow restored by BTMDW-DOC-K-means and that restored by DBN with the planned task flow, shown in Fig.4. It can be seen, task flow restored by BTMDW-DOC-K-means is better than task flow restored by the DBN topic model in two aspects.

Fig.4 shows the task flow restored by BTMDW-DOC-K-means is more closely matched with the planning task flow than that restored by DBN. In the planned task flow, the specific design phase is divided into two tasks, while in the BTMDW-DOC-K-means task flow, it is considered one task. Besides, BTMDW-DOC-K-means extracted a task that does not exist in the planned task flow, paper submission and presentation. Further analysis shows that the project is completed by the university, and this task has existed during the project. It is an extra task that students need to complete outside the traffic wave project.

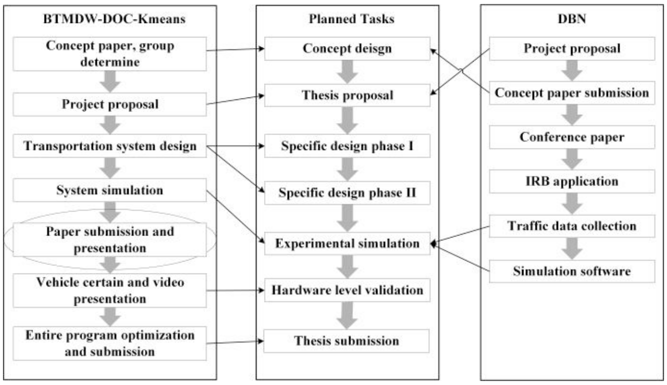

Fig. 4. Task Restored by BTMDW-DOC-K-means and DBN.

Compared to the DBN topic model, BTMDW-DOC-K-means has a lower human intervention. The final tasks extracted by the DBN topic model are selected by humans from 50 topics, which relies on human experience and has great uncertainty due to different understanding. The whole process of the BTMDW-DOC-K-means approach does not depend on human experience directly and will significantly improve its stability.

Besides, the tasks extracted by the DBN topic model are out of order because time information is not considered, which leads to the wrong order of the first and second tasks in Fig.4. In BTMDW-DOC-K-means, the order of tasks depends on the time information of document clusters, which effectively increases the accuracy of restored task flow.

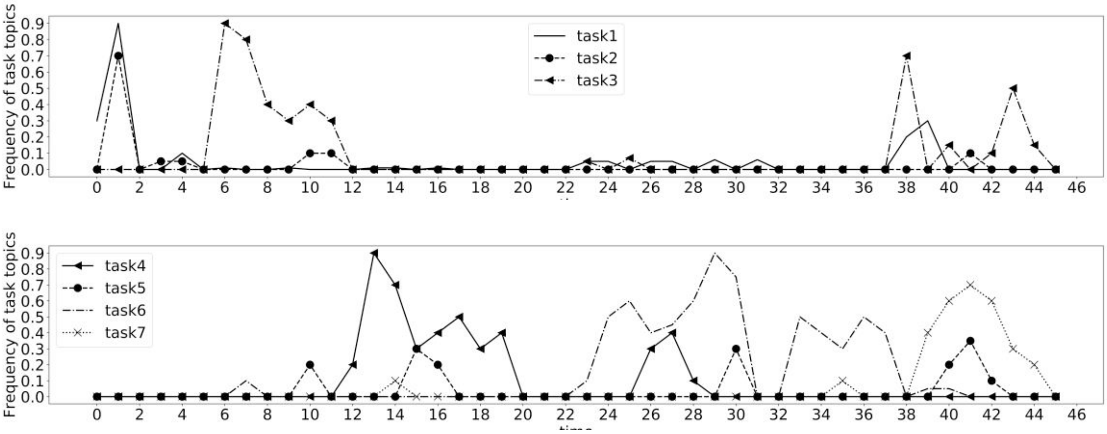

Fig. 5. Occurrence Frequency of Business Tasks.

To find the bottleneck task, we track the timeline of each task and show the design process flow shown in Fig.5 and Fig.6. Fig.5 shows task 1, task 3, task 5 have several peaks, and the second prominent peak occurred during the execution time of task six. During the execution of task 6, some unreasonable places in task1, task3 and task5 were found, and some adjustments or reworks were made.

It can also be seen in Fig.6 that there are three loops in task 6. Besides, completing task6 also takes the longest time according to Fig.5. Considering these two points, task6 can be considered as the bottleneck task of this project. The discovery of bottleneck tasks can provide suggestions to designers that the feasibility of implementation on hardware should be fully considered in the conceptual design stage

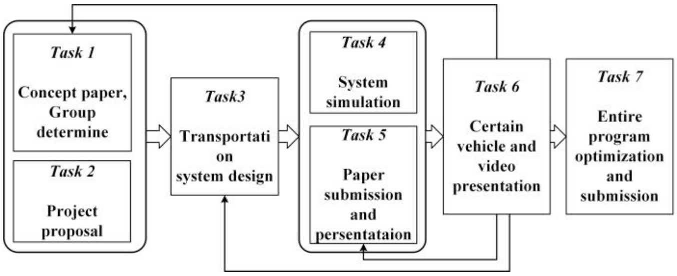
Task 1
Concept paper, Group determine
Task 2
Project proposal
Task3
Transportati on system design
Task 4
System simulation
Task 5
Paper submission and persentataion
Task 6
Certain vehicle and video presentation
Task 7
Entire program optimization and submission

Fig. 6. Design Process Reconstructed by BTMDW-DOC-K-means.

5. CONCLUSION

This paper proposed a novel method called BTMDW-DOC-K-means to automatically identify bottleneck tasks in a business process. We collected the design process data across two years and applied our approach to reconstruct design task flow and identify bottleneck tasks. We then evaluated the usefulness of the algorithm we proposed by comparing it with DBN topic model. The BTMDW-DOC-K-means can extract more coherent keywords related to the topic and give a more suitable representation of documents in different lengths. The fusion-based clustering can extract document features from different aspects and improve performance in the short text, long text, and hybrid text. The significance of this study lies in developing a new approach to identify bottleneck tasks in business processes and improving related algorithms according to the characteristics of business process documents.

REFERENCES

Bengio, Y.. (2009). Learning deep architectures for AI, *Foundations and Trends in Machine Learning.*, 2(1): 1–27.

Blei, D. M., Ng, A. Y., Jordan, M. I.. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research*, 3(5): 993–1022.

Boessenkool H, Wildenbeest J G W, Heemskerk C J M. (2018). A task analysis approach to quantify bottlenecks in task completion time of telemanipulated maintenance, *Fusion Engineering and Design*, 129: 300-308.

Bozorgi Z D, Teinemaa I, Dumas M. (2020). Process mining meets causal machine learning: discovering causal rules from event logs, In *International Conference on Process Mining (ICPM) on*, page 129-136.

Cheng X, Yan X, Lan Y. (2014). Btm: topic modeling over short texts, *IEEE Transactions on Knowledge and Data Engineering*, 26(12): 2928-2941.

Y Liu, H Zhang, C Li, et al. (2012). Workflow simulation for operational decision support using event graph through process mining. *Decision Support Systems, 52(3): 685-697*

Elleuch M, Ismaili O A, Laga N. (2020a). Discovering activities from emails based on pattern discovery approach, In *International Conference on Business Process Management on*, page 29-30.

Elleuch M, Ismaili O A, Laga N. (2020b). Discovery of activities' actor perspective from emails based on speech acts detection, *In International Conference on Process Mining (ICPM) on*, page 73-80.

Gehler, P. V., Holub, A. D., Welling, M.. (2006). The rate adapting poisson model for information retrieval and object recognition, In *The 23rd International Conference on Machine Learning (ICML) on*, page 337–344.

Gupta P, Chaudhary Y, Buettner F. (2019). Document informed neural autoregressive topic models with distributional prior, In *The AAAI Conference on Artificial Intelligence on*, page 6505-6512.

Hinton, G. E., and Salakhutdinov, R.. (2009). Replicated softmax: an undirected topic model, In *The 23rd Annual Conference on Neural Information Processing Systems (NIPS) on*, page 1607–1614.

Huang B, Wang W, Ren S. (2019). A proactive task dispatching method based on future bottleneck prediction for the smart factory, *International Journal of Computer Integrated Manufacturing*, 32(3): 278-293.

Janssenswillen G, Depaire B, Swennen M. (2019). BupaR: enabling reproducible business process analysis, *Knowledge-Based Systems*, 163: 927-930.

Lijun Lan, Y Liu, Feng Lu W. (2017). Learning from the past: uncovering design process models using an enriched process mining, *Journal of Mechanical Design*, 140(4): 041403.

Lijun Lan, Y Liu, Feng Lu W. (2018). Automatic discovery of design task structure using deep belief nets, *Journal of Computing and Information Science in Engineering*, 17(4), 041001.

Li W, Feng Y, Li D. (2016). Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm, *Automatic Control and Computer Sciences*, 50(4): 271-277.

Lu F, Tao R, Du Y. (2019). Deadlock detection-oriented unfolding of unbounded Petri nets, *Information Sciences*, 497: 1-22.

Park S, Lee J, Kim K. (2019). Semi-supervised distributed representations of documents for sentiment analysis, *Neural Networks*, 119: 139-150.

Sun H W, Liu W, Qi L. (2020). A process mining algorithm to mixed multiple-concurrency short-loop structures, *Information Sciences*, 542: 453-475.

Vázquez-Barreiros B, Mucientes M, Lama M. (2016). Enhancing discovered processes with duplicate tasks, *Information Sciences*, 373: 369-387.

Yamin W, Yue H. (2016). Hotspot detection in microblog public opinion based on biterm topic model, *Journal of Intelligence*, 35(11): 119-124.

Yan X, Guo J, Lan Y. (2013). A biterm topic model for short texts, In *The 22nd international conference on World Wide Web on*, page 1445-1456.

Zheng Y, Meng Z, Xu C. (2015). A short-text oriented clustering method for hot topics extraction, *International Journal of Software Engineering and Knowledge Engineering*, 25(03): 453-471.

Zhu F, Wang R, Wang C. (2019). Intelligent workshop bottleneck prediction based on complex network, In *IEEE International Conference on Mechatronics and Automation (ICMA) on,* page 1682-1686.