INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 16, Issue: 2, Month: April, Year: 2021
Article Number: 4101, https://doi.org/10.15837/ijccc.2021.2.4101

Revealing New Technologies in Ocean Engineering Research using Machine Learning

X. Li, Y. Liang, B. Chen, B. He, Y. Jiang

Xin Li, Biqian Chen, Baorun He, Yu Jiang*
1. College of Computer Science and Technology, Jilin University
2. Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, Jilin University
Changchun 130012, China
*Corresponding author: jiangyu2011@jlu.edu.cn

Yanchun Liang
1.College of Computer Science and Technology, Jilin University
2.Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, Jilin University
Changchun 130012, China
3.Zhuhai Sub Laboratory of Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, Zhuhai College of Jilin University
Zhuhai 519041, China

Abstract

On par with aerospace engineering, ocean engineering has caught a lot of attention recently. In this paper we employ machine learning and natural language processing methods to reveal new technologies and research hotspots in the ocean engineering field. Our data collection includes 14 high-impact journals, and the abstracts of almost 30,000 papers published from 2010 to 2019. We employed two topic models, Latent Dirichlet Allocation (LDA) and PhraseLDA. Used independently, the LDA model may lack interpretability and the PhraseLDA result may lose information in the final topics. We hence combined these two models and discovered the research hotspots for each year using affinity propagation clustering and word-cloud-based visualization. The results reveal that several topics such as "wind power" and "ship structure," areas such as the European and Arctic seas, and some common research methods are increasing in popularity. This work consists of data collection, topic modelling, clustering, and visualization, which can help researchers understand the trends and important topics in ocean engineering as well as other fields.

Keywords: Ocean Engineering, Latent Dirichlet Allocation, Machine Learning

1 INTRODUCTION

The ocean occupies 71% of the Earth's surface and 97% of the Earth's water. Ocean engineering is a multidisciplinary research field focusing on solving engineering problems associated

with ocean environment and intelligently explore and harness the ocean's resources. Ocean engineering technology covers a wide range of fields, including ocean power generation, ocean drilling, seawater desalination, offshore oil mining, coastal wind power, sea level detection, ocean material separation, seawater refining, autonomous underwater vehicle, and ocean architecture design [1]-[6]. Ocean engineering is also a technical field that concerns the shipbuilding industry [6], especially ship structural optimization [7] and ship wave resistance [8].

As ocean engineering develops, many research articles have been published. It is increasingly important for researchers to understand the state-of-the-art improvements in ocean engineering, and some researchers have proposed that machine learning methods could be used to determine the topic or main idea of an article to improve research efficiency. An increasing amount of research has focused on the methods of natural language processing to extract popular topics from articles [9]. Moreover, learning from short texts or abstracts has become a critical and complex task [10]. Of these methods, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) [11] have been widely used. NMF [12] works well for topic models and, because it is based on probability distributions, the results are easier to understand. However, NMF does not consider any prior knowledge about the topic probability distribution. LDA addresses this problem, and we hence employ the LDA and PhraseLDA models to capture popular topics and state-of-the-art technologies. The main process used to build our model is as follows. First, we collected the data of 14 ocean engineering journals from 2010 to 2019. Then, we used word clouds to illustrate the topics of each year, and by analyzing the word cloud, we were able to surmise the hotspots and trends in ocean engineering research.

2 RELATED WORK

2.1 LDA TOPIC MODEL

The LDA approach was proposed in 2003 by Blei et al. [13]. It is an unsupervised topic modelling method that represents the topic of each document in the corpus as a probability distribution. LDA is a typical bag-of-words model.

The generation of a model using LDA is as follows: For each document, it extracts one topic from the topic distribution. Then, a word corresponding to the extracted topic is selected from the word distribution. These two steps are iteratively repeated until each word in the documents has been visited.

In LDA, w denotes the N words in a document, D denotes a collection of M documents and N obeys a Poisson distribution. The distribution of topic k over a word is φ_k and $k \in \{1, \ldots, K\}$, and φ_k obeys a Dirichlet distribution with hyperparameter β. The distribution of the mth document over all topics is denoted as θm, where $m \in \{1, \ldots, M\}$ and θm obeys the Dirichlet distribution with hyperparameter α. Further, $z_{m,n}$ denotes the topic of the nth word in the mth document, and $w_{m,n}$ denotes the nth word in the mth document, which is generated by φ_k [13].

The whole process can be expressed simply as follows:

$$p(w|d) = p(w|z) \times p(z|d) \tag{1}$$

where w denotes the word, z represents the topic, and d is the document. Then, the objective function is as follows:

$$p(w, z|\alpha, \beta) = p(w|z, \beta) \times p(z|\alpha) = \prod_{k=1}^{K} \frac{\Delta(\varphi_k + \beta)}{\Delta(\beta)} \prod_{m=1}^{M} \frac{\Delta(\varphi_m + \alpha)}{\alpha)} \tag{2}$$

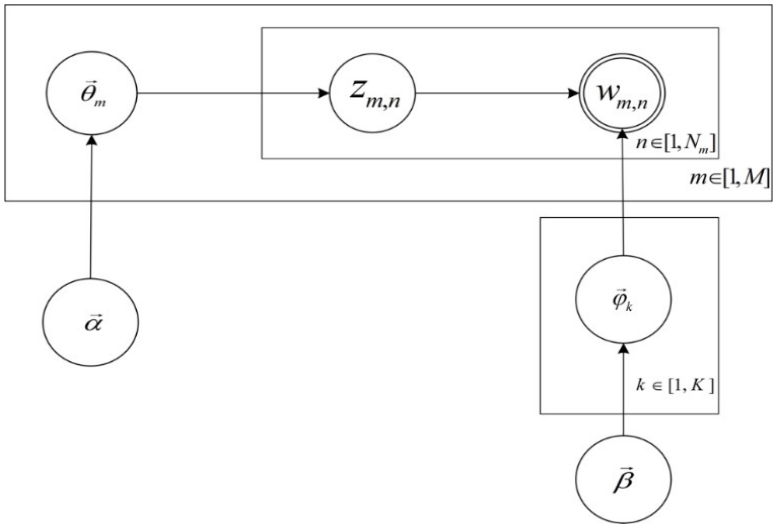

Figure 1: Structure of LDA.

To obtain the LDA solution, we employ a Gibbs sampling algorithm, where the values of α and β are known as priories. Our goal is to obtain the document-topic and topic-word distributions, which are $z_{m,n}$and $w_{m,n}$ respectively. Moreover, we also need to obtain the conditional probability distribution corresponding to each feature dimension for the required target distribution, which is also found by Gibbs sampling of the LDA model. It is formulated as follows:

$$p(z_i = k | \overrightarrow{z}_{\neg} i, \overrightarrow{\omega}) \propto p(z_i = k, \omega_i = t | \overrightarrow{z}_{\neg} i, \overrightarrow{\omega}_{\neg} i) \tag{3}$$

Here, z_i represents the ith word in the corpus, $I = (m, n)$, and i represents the whole corpus except for the ith word.

2.2 PHRASELDA MODEL

The PhraseLDA model is a variant of LDA. We use it as a comparative model to LDA. PhraseLDA was proposed by Kishky et al. in 2014 [14]. The authors noted that most of topic models are based on unigrams, which loses the semantic information in the phrase and influences human interpretation. PhraseLDA is a two-step process. The first step segments the document into single and multi-word phrases, and the second step is the topic model, which is a variant of LDA.

The PhraseLDA uses a bag-of-phrases input instead of the traditional bag-of-words. The first step uses the TopMine framework, which consists of phrase mining, frequent phrase mining, and phrase filtering. To obtain high-quality phrases, the mining process consists of two steps. One is document partitioning, which extracts the candidate phrases and aggregates the phrases' counts. The next step merges words in each document into quality phrases. After phrase mining, frequent phrase mining is used to find phrases with the following two properties: downward closure and data antimonotonicity. All selected phrases must exceed a minimum support threshold. A larger minimum support threshold increases the precision and recall of the model. Then, the model continues to filter phrases, and the key process of this step is a bottom-up merging process. It uses a greedy algorithm to merge single and multi-word phrases guided by best score.

PhraseLDA is a variant of LDA based on the "bag-of-phrases" hypothesis [15]. It uses an undirected graph to model stronger dependences for nearby words. This model selects the gth phrase of the dth document to form a clique, which is the input of a function f and is represented

as follows:

$$p(z, w, \varphi, \theta) = \frac{1}{C} p_{LDA}(z, w, \varphi, \theta) \prod_{d,g} f(c_{d,g}) \tag{4}$$

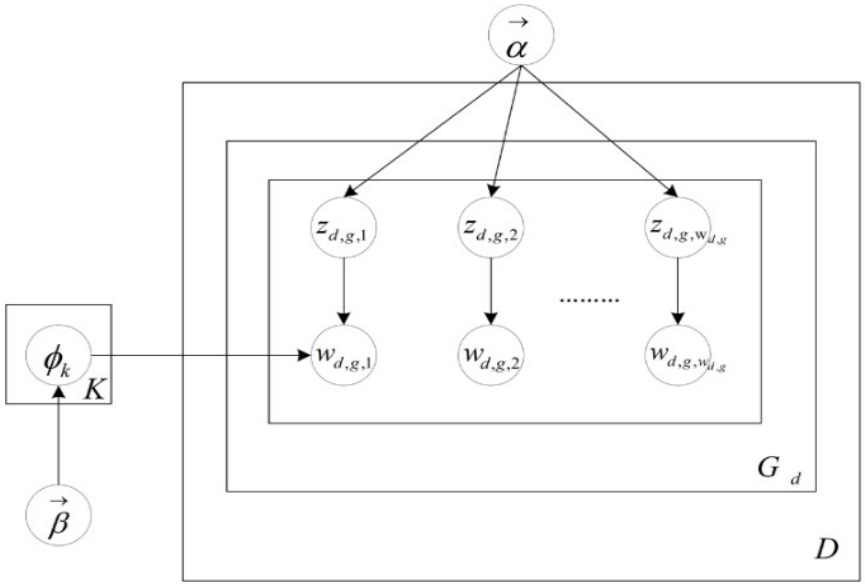

Figure 2: Structure of PhraseLDA.

2.3 AFFINITY PROPAGATION CLUSTERING

Affinity propagation clustering was proposed by Frey and Dueck in 2007. It is an unsupervised method, so the number of cluster centers does not need to be set. All data points are potential cluster centers initially. The model is not sensitive to the initial values of the data, and compared with the k-means method, the squared error is smaller

The key component of the model is its alternate updating of the responsibility matrix and availability matrix. This model uses $s(i,k)$ to denote the similarity between point i and point k. In addition, $r(i,k)$ represents the responsibility value between i and k: this value represents the probability that point i selects k as its cluster center. Finally, $a(i,k)$ indicates that center k sees i as an attractive point to have in its cluster. The preference value and damping factor are two factors that substantially affect the result. The preference refers to the degree to which point i prefers itself as the cluster center. In general it is set to the median of the similarity value [16]. The damping factor is λ.

The main processes of affinity propagation is as follows. The update of $r(i,k)$ is

$$r(i,k) \leftarrow s(i,k) - \max_{k" s.t. k" \neq k} \{a(i,k") + s(i,k")\} \tag{5}$$

when $i = k$, $r(k,k)$ indicates whether a point is suitable to be its own center. The update of $a(i,k)$ is

$$a(i,k) \leftarrow \max \left\{ 0, r(k,k) + \sum_{i" s.t. i" \notin \{i,k\}} \max\{0, r(i",k)\} \right\} \tag{6}$$

$$a(k,k) \leftarrow \sum_{i" s.t. i" \neq k} \max\{0, r(i",k)\} \tag{7}$$

After each iteration, the two messages update according the following functions:

$$r_{new}(i,k) = \lambda * r_{old}(i,k) + (1-\lambda) * r(i,k) \tag{8}$$

$$a_{new}(i,k) = \lambda * a_{old}(i,k) + (1-\lambda) * a(i,k) \tag{9}$$

3 OCEAN ENGINEERING HOTSPOT DETECTION MODEL

3.1 WEB CRAWLER FRAMEWORK

To collect scientific manuscripts, we constructed a web crawler. Before crawling information online, we consulted experts about which journals to crawl and identified 14 journals as data sources. Then, we analyzed the websites of these journals. We need the URLs of the articles are essential to crawl the data. However, we found that different websites have different URL structures. For some websites, we could not collect the article URLs directly. Therefore, we identified the URLs of the issues first and whether these websites include secondary directories. If a URL could not be accessed, we checked whether it was mistyped. Then, we analyzed the website structure. If we could collect an article URL directly, we would extract four types of information of the current articles (title, author list, publication date, and abstract). Otherwise, we stored the hierarchical structure of the journal's website. Then, we obtained the article URLs to extract the article information.

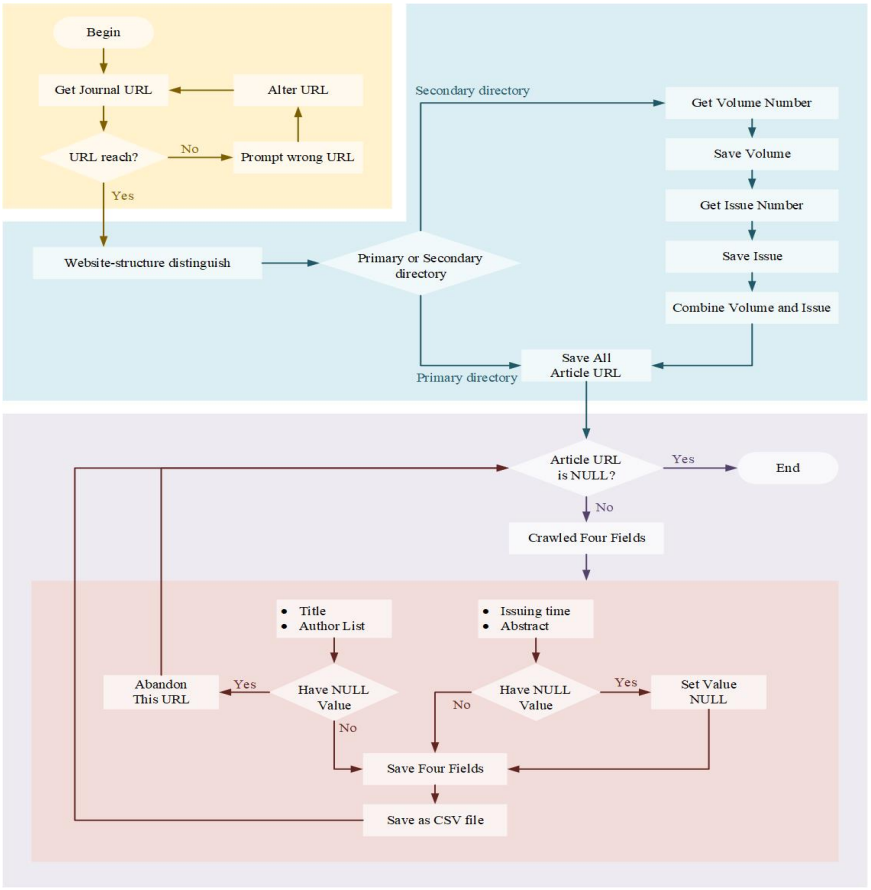

Figure 3: Web crawler framework.

3.2 HOTSPOT AND TREND ANALYSIS

Our model consists of three steps. The first is data collection (using the web crawler) along with data cleaning and pretreatment. These processes are essential for improving the quality of data. The second step is topic mining. In this step, we used LDA and PhraseLDA to mine the candidate topics. The next step is topic clustering, and the hotspots are summarized accordingly. And then affinity propagation clustering was used to cluster the topic identify the hotspots of ocean engineering for each year. In addition, we used word clouds to visualize the results and predict research trends.

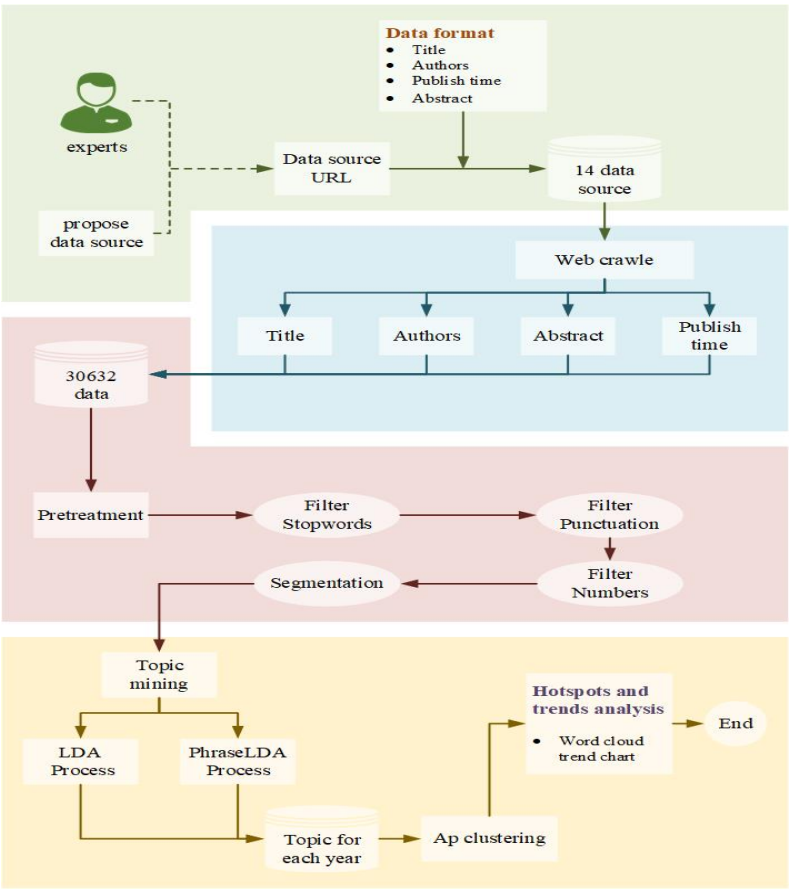

Figure 4: Hotspot discovery and trend analysis.

In our experiments, we used the Gensim package for preprocessing. We transformed upper case letters into lower case ones, filtered out punctuation and numbers, and performed segmentation. Moreover, we used perplexity(P) to find the best number of topics. Perplexity is computed as follows:

$$Perplexity = e^{\frac{-\sum \log(p(w))}{N}} \tag{10}$$

where N denotes the counts of all the words in the test set, and $p(w)$ is the probability of each word in the test set. The perplexity decreases as the numbers of topics increase, but we cannot allow the number of topics to infinitely increase. Therefore, we chose a topic number that is

Table 1: COLLECTED DATA

Journal title	Number of articles	Final number of articles after preprocessing
Engineering Structures	5,899	5,122
Journal of Computational Physics	6116	5299
ACHhE Journal	3416	2735
Computers and Fluids	3081	2756
Journal of Applied Physics	4557	4545
Science China-Technological Sciences	2032	1987
Ocean Engineering	1741	3912
Archive of Applied Mechanics	1041	1028
Journal of Hydrodynamics	756	680
China Ocean Engineering	439	426
Annual Review of Fluid Mechanics	449	209
Journal of Fluids and Structures	334	1403
Journal of Marine Science and Technology-Taiwan	358	329
American Institute of Aeronautics and Astronautics Journal	571	201

less than the number of input documents to obtain the best perplexity. We set the number of iterations $i = 500$ and set hyper parameters $\alpha = 0.25$ and $beta = 0.01$. The output of LDA is a matrix of 30 columns that indicates the 30 words that are most suitable for representing a topic, which is represented by each row. For affinity propagation clustering, we set the preference to the median of s, and the damping factor to 0.9. For PhraseLDA, we used the version by El-Kishky [14] and fine-tuned the parameters. We let the threshold = 5 and the maximum phrase length = 3. According to the results of a comparative test, we chose a suitable number of topics for each year and let the number of words in each topic range from 10 to 20.

4 EXPERIMENT

4.1 DATASET

To collect relevant and high-quality data, we considered the ranking of journals in ocean engineering from the Scientific Journal Rankings website [17]. We also consulted experts in ocean engineering to confirm the relevance of these journals. We collected about 40 journals from the Scientific Journal Rankings website and kept 14 final data source websites after filtering them according to impact factor and relevance.

4.2 RESULTS

To illustrate the hotspots, we introduce word clouds. At the beginning, we divided all data into ten sets according to the publication year of each article. Then, we applied the LDA and PhraseLDA models for each set to extract popular topics. The results are presented as word clouds in Fig. 5 and 6, respectively.

Several conclusions can be inferred from the two figures. One is that the numbers of words or phrases increase annually for all years except for 2019. We find that the amount of crawled data increases more than six times (from 895 in 2010 to 5788 in 2018). This indicates that there

is increased interest in ocean engineering research. Moreover, the number of words produced by the LDA model is larger than the number of phrases produced by PhraseLDA when they reach convergence. Moreover, we find that LDA produced more topics than PhraseLDA. This is probably because PhraseLDA mines phrases before generating topics, as a consequence of which, it reduces the amount of information.

Although LDA cloud has a higher word density, we find that LDA yields less information than PhraseLDA. Moreover, it is very hard to infer the topic from the word clouds that represent them. For example, in the word clouds for 2013, "fiber" in Figure 5(d) is less informative than "fiber orientation" in Figure 6(d). The results also include some interesting aspects such as mathematical and methodology words (e.g., "numerical simulations," "Boltzmann," and "Coriolis"), which could provide guidance for future study. The PhraseLDA results are full of phrases that are easier to explain for most cases. The expressive ability of PhraseLDA is much better than that of LDA. We find mathematical and methodology phrases as well as domain-related phrases. Besides these specific topics, we also find several existing research hotspots such as "wind tunnel", "control system", and "magnetic field" However, for the PhraseLDA, similar or the same words exist in more than one phrase. Phrases with the same word may share similar semantics, but more diversity is demanded for advanced technology discovery in ocean engineering.

When we analyze the topic results, we find that there are several words like "global" and "area" that might present some research hotspots in particular areas of the ocean or indicate a country or location dedicated to the development of the ocean. The idea inspired us to look deeper into the issue. First, we mined the whole dataset again and used a neural network based named entity recognition model [18], to extract the region words for each year. Terms like "the Black Sea" and "Arabian Sea" represent a certain sea area or region. The results are presented in Figure 7, and we find that the words with the highest frequency are "Europe", "Arctic", and "the South China Sea". This suggests that there are some research concerns in these areas. We also mined locations, and the results are presented in Figure 8, where we find that "China" is the most frequently used word. This indicates that China attracts a large amount of attention to ocean development, or it could indicate that some Chinese researchers have a high impact in the ocean engineering field.

Specially, in the LDA results, we find some research methods such as "Boltzmann", "Navier-Stokes", and "Reynolds". These terms belong to the field of ocean engineering or fluid mechanics. We also find that the word "nanowire" indicates a new material applied in marine power. In the PhraseLDA results, we find research method topics such as "Monte Carlo" and "sensitivity analysis". We also find "offshore wind turbines" appearing in 2016, which supports the trend in industrial applications. In 2015, the increase in the number of new wind turbines peaked in Europe, as shown in Figure 9, so the corresponding results in 2016 would be expected considering the publication cycle.

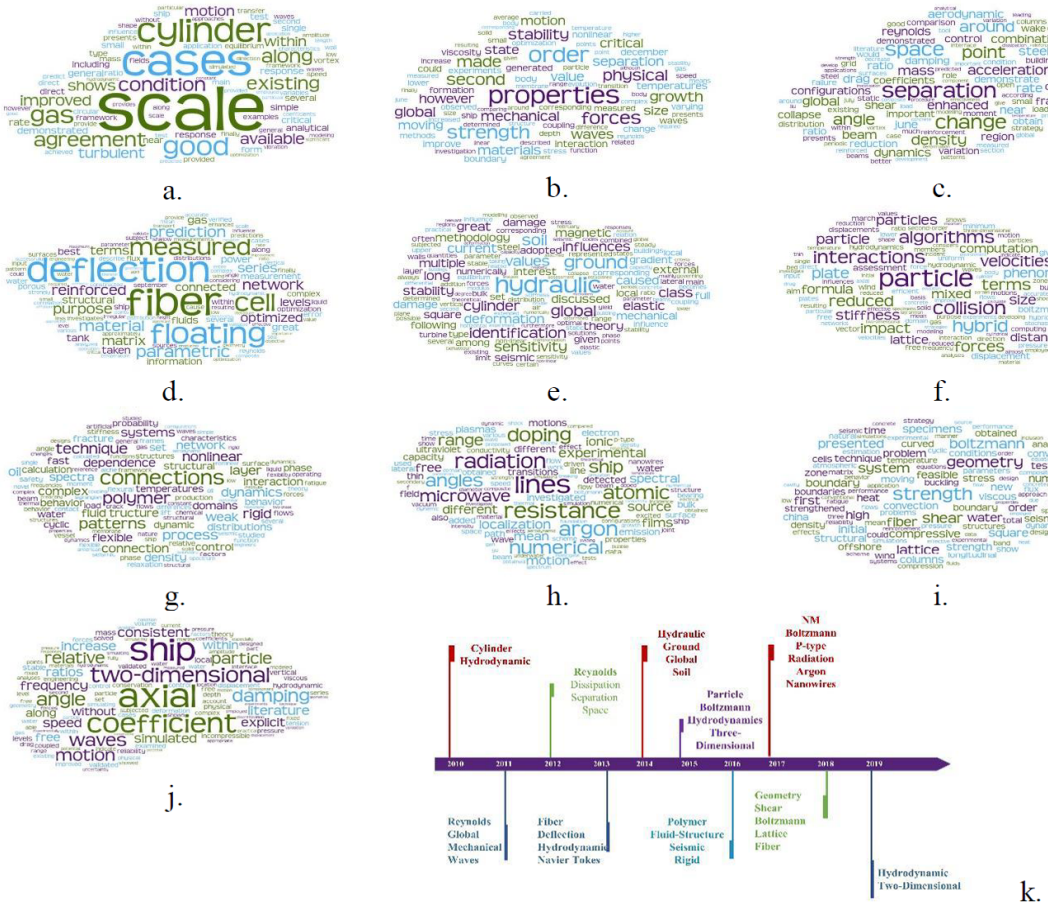

Figure 5: Hotspots in ocean engineering research from 2010 to 2019 obtained using classical LDA.

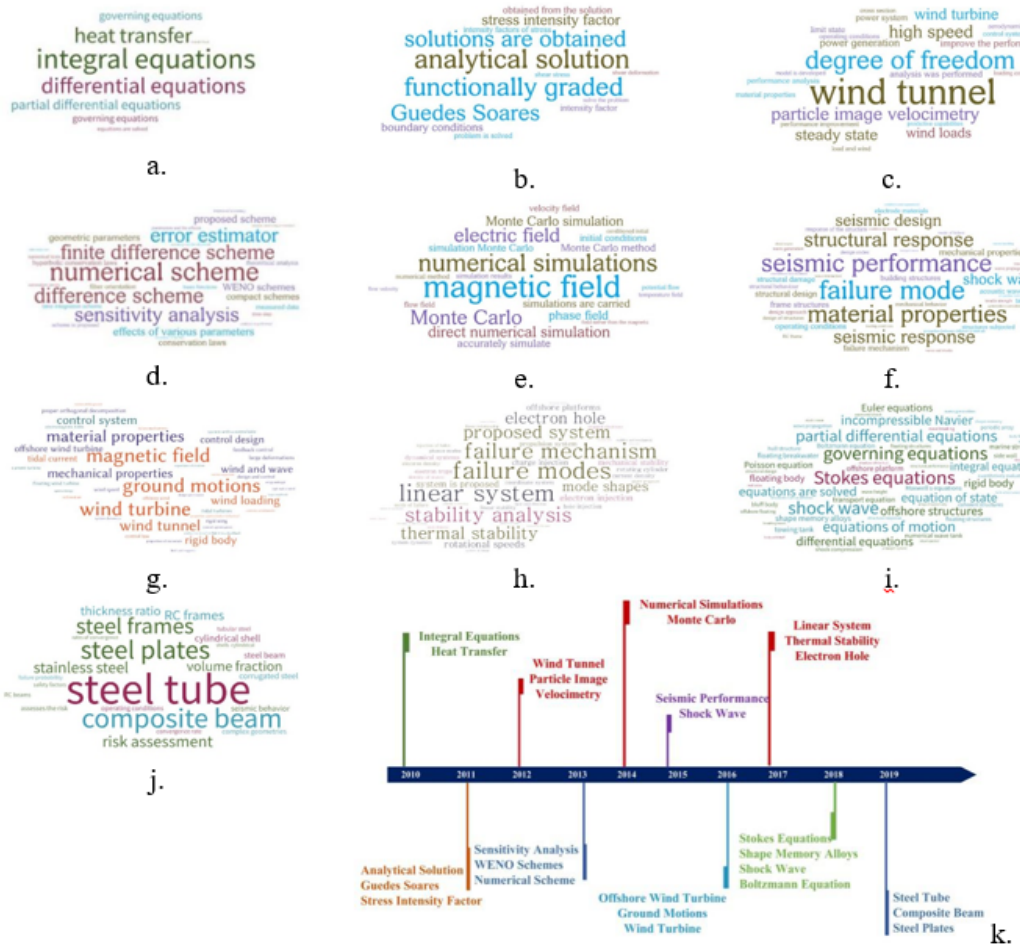

Figure 6: Hotspots in ocean engineering research from 2010 to 2019 obtained using PhraseLDA.

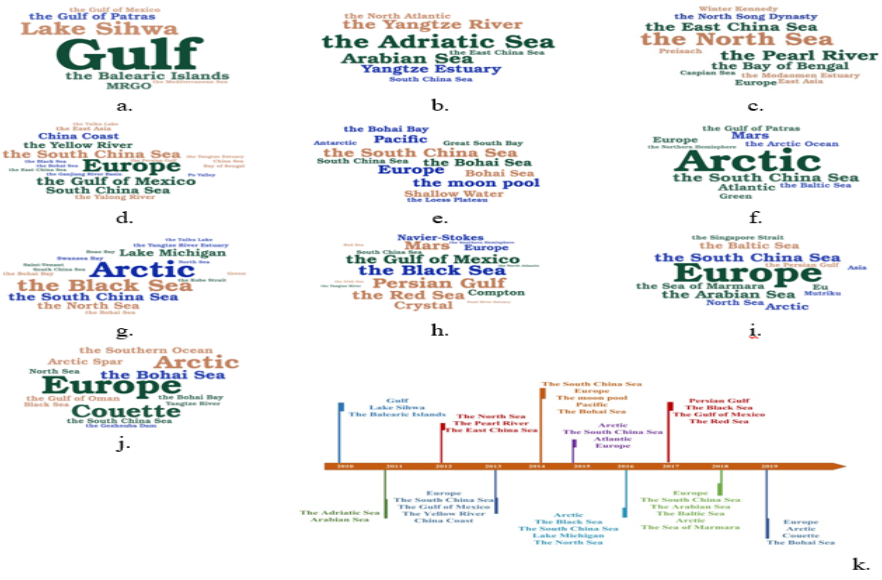

Figure 7: Popular areas in ocean engineering research.

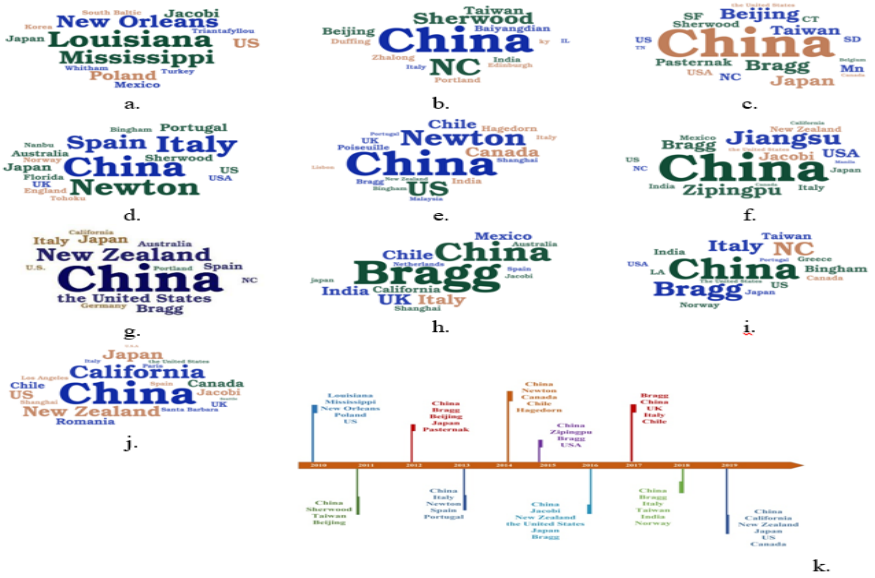

Figure 8: Popular cities, states, and countries for ocean engineering research.

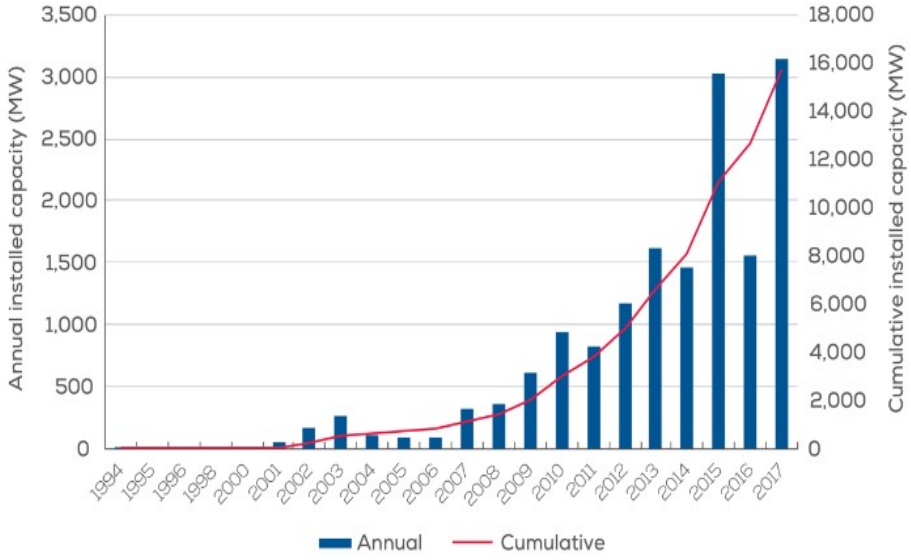

Figure 9: European offshore wind power [19].

5 CONCLUSION

In this study, we used natural language processing and machine learning to reveal new technologies and research hotspots in the ocean engineering field. Moreover, we proposed a new framework that consists of data collection, topic model, clustering algorithm, and visualization. Our data collection contains 14 high-impact journals, and the abstracts of almost 30,000 papers from 2010 to 2019. To comprehensively explore the problem, we employed two topic models: LDA and PhraseLDA. These two models are not only for comparison, but also for complementation. We found that when they are used independently, the LDA model lacks interpretability and the PhraseLDA result is not satisfactory because phrase mining leads to information loss

and the complete topic information cannot be extracted. We hence combined these two models and discovered the research hotspots for each year. The results reveal that several fields such as "wind power" and "ship structure" have become hotspots during the past few years. We also find the most researchers focus on the European and Arctic sea areas. Moreover, we also revealed several common research methods. In the future, we are going to explore more journals to reveal more knowledge in ocean engineering.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61972174, Grant 62072211, the Science and Technology Planning Project of Guangdong Province under Grant 2020A0505100018, the Guangdong Premier Key-Discipline Enhancement Scheme under Grant 2016GDYSZDXK036, and the Guangdong Key-Project for Applied Fundamental Research under Grant 2018KZDXM076.

References

[1] H. Qin, C. Wang, Y. Jiang, Z. C. Deng, and W. Zhang. (2018); Trend prediction of the 3D thermocline"s lateral boundary based on the SVR method, *EURASIP Journal on Wireless Communications and Networking*, 1, pp.252, 2018.

[2] Y. Jiang, M. Zhao, C. Hu, L. He, H. Bai, and J. Wang. (2019) A parallel FP-growth algorithm on World Ocean Atlas data with multi-core CPU, *The journal of Supercomputing*, 2, 732-745, 2019.

[3] M. H. Zhao, C. Q. Hu, F.L. Wei, K. Wang, C. Wang, Y. Jiang. (2019) Real-Time Underwater Image Recognition with FPGA Embedded System for Convolutional Neural Network, *Sensors*, 2, pp.350 2019.

[4] Y. Jiang, Y. Gou, T. Zhang, K. Wang, C. Hu. (2017) A machine learning approach to argo data analysis in a thermocline, *Sensors*,10, pp.2225, 2017.

[5] H. D. Qin, H. Chen, and Y.C. Sun. (2019) Distributed finite-time fault-tolerant containment control for multiple Ocean Bottom Flying Nodes, *Journal of the Franklin Institute*, doi:10.1016/j.jfranklin.2019.05.034, 2019

[6] Y. Jiang, T. Zhang, Y. Gou, L. He, H. Bai, and C. Hu.(2018) High-resolution temperature and salinity model analysis using support vector regression, *J. Ambient Intell. Hum. Comput*, 1-9, 2018.

[7] H. D. Qin, H. Chen, Y.C. Sun, L.L. Chen. (2019) Distributed finite-time fault-tolerant containment control for multiple ocean Bottom Flying node systems with error constraints, *Ocean Engineering*, doi: 10.1016 /j.oceaneng.2019.106341, 2019

[8] P. Kujala et al.(2019) Review of risk-based design for ice-class ships, *Mar. Struct.*, 63, 181-195, 2019

[9] O. Hizir, M. Kim, O. Turan, A. Day, A. Incecik, and Y. Lee. (2019) Numerical studies on non-linearity of added resistance and ship motions of KVLCC2 in short and long waves, *Int. J. Nav. Archit. Ocean Eng.*, 1, 143-153, Jan. 2019.

[10] A. M. Cohen (2005) A survey of current work in biomedical text mining, *Brief. Bioinform.*, 1, 57-71, 2005.

[11] M. Rei. (2017) Semi-supervised Multitask Learning for Sequence Labeling, *in Proc. 55th ACL 2017*, 2121-2130, 2017.

[12] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin (2019) Experimental explorations on short text topic mining between LDA and NMF based schemes, *Knowl.-Based Syst.*, 163, 1-13, 2019.

[13] D. D. Lee and H. S. Seung(2001) Algorithms for non-negative matrix factorization, *Adv. Neural Inf. Process. Syst.*, 556-562, 2001.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. (2003) Latent Dirichlet allocation, *J. Mach. Learn. Res.*, 3, 993-1022, 2003.

[15] A. El-Kishky, Y. Song, C. Wang, C. Voss, and J. Han (2014) Scalable topical phrase mining from text corpora, *in Proc. VLDB endowment*, 3, 305-316, 2014.

[16] Y.K Tang, X.L Mao, and H.Y Huang. (2016) Labeled phrase latent Dirichlet allocation, *in Proc. WISE 2016*, 525-536, 2016

[17] B. J. Frey and D. Dueck. (2017) Clustering by passing messages between data points, *Science*, 5814, 972-976, Feb. 2007.

[18] [8] Scientific Journal Rankings website, 2019. https://www.scimagojr.com/journalrank.php

[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. (2016) Neural architectures for named entity recognition, *in Proc. NAACL-HLT*, 260-270, 2016.

[20] WindEurope, "WindEurope-Annual-Statistics-2018," 2018. https://windeurope.org/.

Cite this paper as:
Li, X.; Liang, Y.; Chen B.; He B.; Jiang, Y. (2021). Revealing New Technologies in Ocean Engineering Research using Machine Learning, *International Journal of Computers Communications & Control*, 16(2), 4101, 2021.
https://doi.org/10.15837/ijccc.2021.2.4101