Hindawi
Scientific Programming
Volume 2022, Article ID 9497783, 11 pages
https://doi.org/10.1155/2022/9497783

Research Article

Construction and Application of a Data-Driven Abstract Extraction Model for English Text

Hui Peng

Wuhan Huaxia University of Technology, Wuhan, Hubei 430000, China

Correspondence should be addressed to Hui Peng; hx0911_ph@hxut.edu.cn

Received 11 January 2022; Revised 18 February 2022; Accepted 24 February 2022; Published 23 March 2022

Academic Editor: Sheng Bin

In this paper, a single English text is taken as the research object, and the automatic extraction method of text summary is studied using data-driven method. This paper takes a single text as the research object, establishes the connection relationship between article sentences, and proposes a method of automatic extraction of text summary based on graph model and topic model. The method combines the text graph model, complex network theory, and LDA topic model to construct a sentence synthesis scoring function to calculate the text single-sentence weights and output the sentences within the text threshold in descending order as text summaries. The algorithm improves the readability of the text summary while providing enough information for the text summary. In this paper, we propose a BERT-based topic-aware text summarization model based on a neural topic model. The approach uses the potential topic embedding representation encoded by the neural topic model to match with the embedding representation of BERT to guide topic generation to meet the requirements of semantic representation of text and explores topic inference and summary generation jointly in an end-to-end manner through the transformer architecture to capture semantic features while modelling long-range dependencies by a self-attentive mechanism. In this paper, we propose improvements based on pretrained models on both extractive and generative algorithms, making them enhanced for global information memory. Combining the advantages of both algorithms, a new joint model is proposed, which makes it possible to generate summaries that are more consistent with the original topic and have a reduced repetition rate for evenly distributed article information. Comparative experiments were conducted on several datasets and small uniformly distributed private datasets were constructed. In several comparative experiments, the evaluation metrics were improved by up to 2.5 percentage points, proving the effectiveness of the method, and a prototype system for an automatic abstract generation was built to demonstrate the results.

1. Introduction

The value of the text is not in the static data but in the value of the data and information that comes from text comprehension and transmission. In recent years, there has been a growing demand for automated processing of large amounts of text, rather than manual annotation, which has forced the need for machines to be trained to learn how humans process text and understand communication [1]. Natural language processing exists for machines that can better mimic human processing of natural language, to be able to perform tasks such as automated voice conversations, automated text writing, and other tasks on large data, as smart as the human brain. In this era of big data, where labour costs are extremely expensive, natural language processing technology can achieve a large amount of information and value from the text, becoming one of the important technologies that will allow humans and machines to communicate without barriers in the future [2]. TFIDF improves the insufficiency of word frequency statistics method. In addition to considering word frequency, it also calculates the inverse document frequency of words. The basic idea is that if a word appears in most of the articles in the corpus, even if the word frequency of the word is high, its TFIDF value is not necessarily high.

Although artificial intelligence has made rapid development in various fields in recent years and computers are closer to the human brain than any human era, computers are not human brains, and they cannot understand the meaning and generate their cognition accurately after

reading some relevant texts as humans do, but they can only process documents through statistics, machine learning, simple reasoning machines, and elementary memory mechanisms [3]. They can only extract or simply "think-process" the document to compose the final summary of the article through statistical, machine learning, simple inference machines, as well as elementary memory mechanisms. The model in this paper is more accurate for the contextual semantic acquisition of long texts, and the ability to rely on long distances is improved. When the input text is short, it is found that the evaluation index results of the pure Transformer model and the PGEN model are similar, indicating that the simple Transformer model has a strong ability to process short texts, and the generation ability can be comparable to that of the LSTM network with the addition of attention mechanism. However, we expect the text summary to be a "deep understanding" of the text, and the computer is not able to "understand" the real meaning of the document. Most of the current research on automatic text summarization tends to extract the sentences that express the core meaning of the text from the original text so that they contain as much information as possible about the text [4]. However, no matter which sentences are extracted from the document, they cannot fully express the main meaning of the text. With the recent technological breakthroughs and innovations in natural language processing tasks by neural network sequence models and distributed representation learning, text summarization and its applications have received increasing attention from researchers.

In the era of social networking, the rapid development of data mining in information retrieval and natural language processing has made automatic text summarization tasks necessary, and how to effectively process and utilize text resources has become a hot research topic [5]. The text summarization task aims at converting text into a summary containing key information. Today's automatic text summarization methods are mainly classified into extractive and generative models. Although these models have strong coding capabilities, they still fail to address the problems of long text dependencies and semantic inaccuracies. Therefore, in this paper, an in-depth study is done to further address the major problem that the generated summaries do not match the source text facts [6].

2. Related Works

The term "data-driven" first came from the field of computer science; when we construct mathematical models that often cannot be solved by accurate and real (generally real principles are simple and accurate) methods, we also construct approximate models to approximate the real situation with a large amount of data refinement based on previous historical data [7], which were derived by data-driven control model. Heldens et al. proposed model-driven data reengineering, model transformation MDE tools for creating metamodels, and model transformation languages. Bernhard Hohmann proposed a GML-based modelling language to generate parameter-driven extraction models [8]. In foreign countries, data-driven approach has gradually switched from data conversion and reengineering, which is generally used for computers, to parametric design and driving of model construction. Xu and Dang of Northeastern University, in "Simulation Research on Data-Driven Modelling Methods," summarized the BP neural network-based model established by TE data drive of the combined heaters site [9]. Xu et al. of Duke University analysed the two-way link between Revit Structure and Robot Structural Analysis and compared the analysis with the calculation results of PKPM [10].

Automatic text summarization tasks have received increasing attention as an important branch of natural language processing tasks. In terms of content, automatic summarization is divided into single-document summarization and multidocument summarization. Methodologically, it is divided into extractive and generative summarization [11]. Topic modelling is one of the powerful tools for text mining, which can mine potential connections between data, and between data and text through a priori knowledge of the text. Topic modelling can be used to its greatest advantage in dealing with source texts of discrete data. These models use Gibb's sampling, nonnegative matrix decomposition, variational inference, and other machine learning algorithms to infer hidden topic information from the feature text space, especially for high-dimensional and sparse feature texts [12]. The probabilistic topic model was born, which extracts the topic words and their probabilistic combinations that can express the text topic from the massive text and largely dissects the document semantics to classify or cluster the text on a deeper level. Early probabilistic topic models represent PLSA and the widely used LDA model, which has attracted increased researchers to improve and apply all aspects of topic models from model assumptions, inference of parameters, and number of topics to supervision. Nadeem et al. have used LDA models to label the topics of source texts and used formal concept analysis to construct structures and so forth. Rajendra et al. have proposed a heuristic approach to ensure that the generated texts contain the necessary compositional information of the original documents of the corpus through a potential Dirichlet assignment technique to match the optimal number of topics of the source texts [13]. In addition, some studies have combined a two-level topic model based on the Pinball Allocation Model (PAM) with the Text Rank algorithm to accomplish topic text summarization. However, these traditional long text topic modelling algorithms based on word cooccurrence have great limitations, and the problem of limited information and vocabulary in the text is not well solved [14].

3. Data-Driven Model Constructions for English Text Summary Extraction

3.1. Data-Driven Model Design. The context of data-driven is graphical elements, and BIM models are represented by three-dimensional geometry containing various types of information. "Data-driven" means that the control starts from data and ends with data, in a "closed-loop" way, turning a bunch of numbers into digital models. Specifically,

it is the process of creating, adjusting, judging, and optimizing data based on online and offline data from controlled systems. When using the data-driven modelling approach, data is used as the characteristic parameters of the physical object to control and generate the model, and changing the values of the characteristic parameters can also be reflected in the 3D model [15]. Model information is the core of the BIM parametric model, and the parametric model contains the characteristic parameters of all components and the parameters of the interaction relationship between components. The flow of data is the key of data-driven, and this paper is mainly based on Revit software, the means of secondary development of Revit, and then the study of Revit model-driven modelling. The model established by the data-driven method in this paper is based on the model created by the 11G101 flat method atlas, which is theoretically consistent with the engineering quantity of our domestic cost calculation software, so the comparison of the calculation of steel bars is carried out. By assigning the defined parameters to drive Revit internal objects, different parametric models can be obtained according to different parameter settings. The main process is shown in Figure 1.

With the advent of the era of big data in transportation, there is an increasing concern about data-driven methods, and several data-driven methods based on them have been proposed, which can be broadly classified into two categories [16]. One category is data analysis models based on mathematical statistics, mainly including historical averaging, linear regression models, time series analysis methods, and Kalman filtering. The other category is artificial intelligence-based data mining models, which mainly include artificial neural networks, support vector machines, nonparametric regression models, wavelet analysis models, and deep learning models that have received increasing attention in recent years.

3.1.1. Historical Average Model. The basic principle of Historical Average Model (HAM) is

$$V(\text{new}) \geq (\alpha + 1)V - V(\text{old}), \tag{1}$$

where V(new) is the traffic flow of the road section in a certain time interval in the future, V(old) is the traffic flow of the road section in a certain time interval in history, V is the traffic flow of the road section in the current time interval, and α is the smoothing factor.

The conventional polynomial chaos expansion method using Nataf transformation in dealing with correlations has the potential to cause probability and risk underestimation, so this section uses a data-driven arbitrary polynomial chaos expansion method for stochastic tidal analysis. If the state variable ϖ of the system (e.g., the node voltage amplitude in the tidal equation) is represented, the idea of the polynomial chaos expansion method is to approximate the state variable by a truncated sequence of polynomial orthogonal bases consisting of random variables X (e.g., node power, etc.). This method can not only meet the task of extracting multisentence abstracts but also be extended to related tasks such as news headline generation. In addition, the model proposed in this paper is not limited to the field of extracting abstracts of articles as it can also be extended to the field of extracting abstracts from other media.

$$\varpi = \sum_{i=1} c_{k_i} p^i(X), \tag{2}$$

where k is a constant corresponding to the polynomial basis $p^i(X)$.

In traditional methods such as the generalized polynomial chaos expansion method, only the orthogonal basis of the independent univariate probability space (e.g., Gaussian distribution, etc.) is considered, so the original correlated multidimensional probability space needs to be removed from correlation by the Nataf transformation before the generalized polynomial chaos expansion method can be used [17]. In the framework of the generalized polynomial chaos expansion method, the orthogonal basis in the multidimensional probability space is the tensor product of the orthogonal bases of the independent univariate probability space.

Another problem when using AVI detector data for travel time estimation is the identification of noisy data versus anomalous data. Unlike GPS floating vehicle data, this problem can still occur even when the sample of data collection vehicles is large enough. Exceptionally long versus exceptionally short travel time samples should be removed from the database to obtain more reliable travel time estimates. However, identifying and processing invalid data from the data is not always so simple and intuitive. Therefore, several methods have been proposed for the identification of invalid data, such as travel time data due to stopovers, duplicate data, and data where the travel speed exceeds the speed limit value. Most of the existing studies use only one type of detector data for travel time estimation, and one data source contains a limited amount of traffic information. To address this problem, some travel time estimation methods that fuse different types of sensor data have been proposed to improve the accuracy of travel time estimation and reduce the cost of data collection, as shown in Figure 2 for a data-driven model. Firstly, the dependency syntax analysis is carried out on the text according to the dependency syntax analysis theory, and the text graph model is constructed with sentences as nodes in combination with the text graph model. The model finally uses the comprehensive scoring function of the sentence to calculate the weight of the sentence in the text and outputs the sentence within the text threshold range as the text summary.

Although the concept of "data fusion" has been described in the literature in several dimensions, the "data fusion" mentioned here is mainly concerned with the fusion of data collected by different types of traffic sensors. This paper considers two completely different types of data. In this paper, we consider two completely different types of data fusion. The first one is fusion at the data level, which is the most direct way of data fusion [18]. It constructs only one model and integrates the data collected by multiple types of sensors as the input data to the model after fusion. The other is decision-level fusion, which constructs separate models with different types of sensor data and then uses various

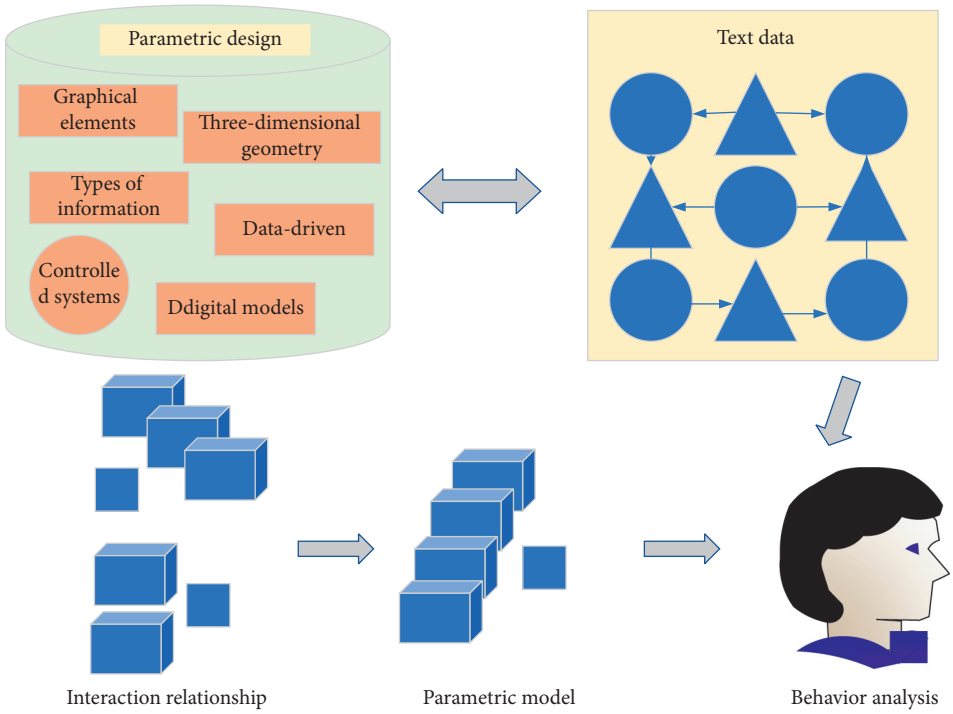

Figure 1: Schematic diagram of the data-driven process.

methods at the decision level to fuse the estimates (i.e., outputs) from multiple models to give the final travel time estimates. Another approach to fusing multiple sensor data is the state-space model. These models consist of two equations that describe the evolution of unobservable state variables and are widely used for the estimation and prediction of time series. The state equation theoretically defines the evolution of the system, while the measurement equation describes the relationship between the observed variables and the true value of the unobservable state. The Kalman filter model is widely used to solve these dynamic state-space models. It is a recursive method and different variants of these filter models exist due to the different degrees of linearity of the equations involved. In the existing studies, different methods for travel time estimation based on state-space models have been proposed, all of which incorporate data from different types of sensors as input to the model.

Finally, the data-driven arbitrary polynomial chaos expansion methods of different orders are compared. As shown in Figure 3, the probability density function of the magnitude on node 5 is given, and the results are obtained from two kinds of stochastic tide analysis methods, namely, Monte Carlo simulation method (MCM) as a comparison criterion, the 2nd-order according-driven arbitrary polynomial chaos expansion (2nd-order APC), and the 3rd-order according-driven arbitrary polynomial chaos expansion (3rd-order APC). The 3rd-order according-driven arbitrary polynomial chaos expansion approximates the probability density function obtained from the Monte Carlo simulation slightly better than the 2nd-order according-driven arbitrary polynomial chaos expansion because the use of higher-order according-driven arbitrary polynomial chaos expansion implies the use of higher-order polynomial basis functions in the polynomial approximation, and more polynomial basis functions lead to a closer approximation to the simulation results. However, the higher-order according-driven arbitrary polynomial chaos expansion method introduces more coefficients to be determined, which also makes the polynomial order of the deterministic system of equations obtained after the stochastic Gallatin integration method higher and more difficult to solve, which brings some difficulties to practical applications.

In summary, although the stochastic Gal Liukin-based method can find the coefficients of the polynomial fit more accurately, it is difficult to be applied to large-scale systems, so the noninvasive data-driven arbitrary polynomial chaos method should be carried out in the subsequent research. A collocation point method is a representative form of non-invasive data-driven arbitrary polynomial chaos expansion method, which only requires computational analysis of the tidal equation at a finite number of sampling points and then uses regression analysis to calculate the coefficients of the polynomial basis and thus has the advantages of small computational effort and low time consumption, but the accuracy is slightly inferior to that of the embedded analysis method based on the stochastic Algonquin integration method.

3.2. English Text Summary Extraction Model Construction. The bag-of-words model does not consider the order relationship of words in a document, and the continuous bag-of-words model only does not consider order information

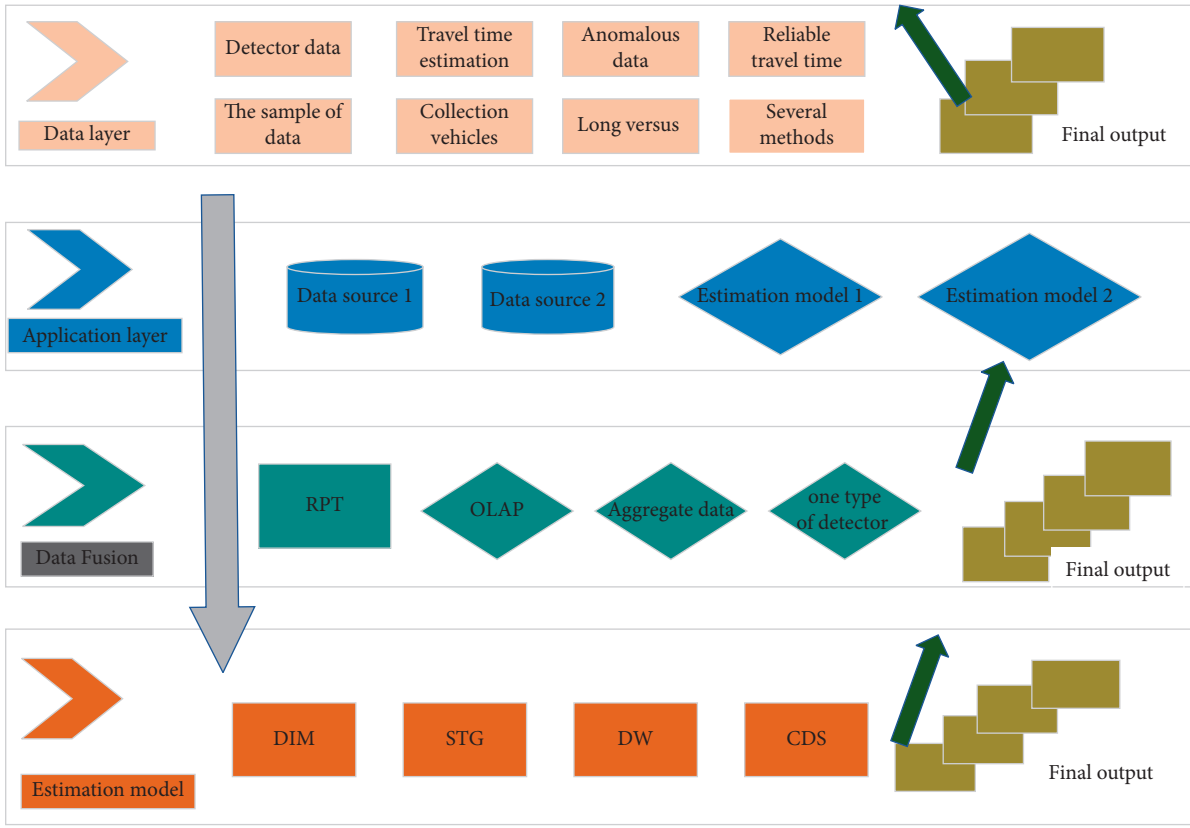

Figure 2: Data-driven English text summarization extraction process.

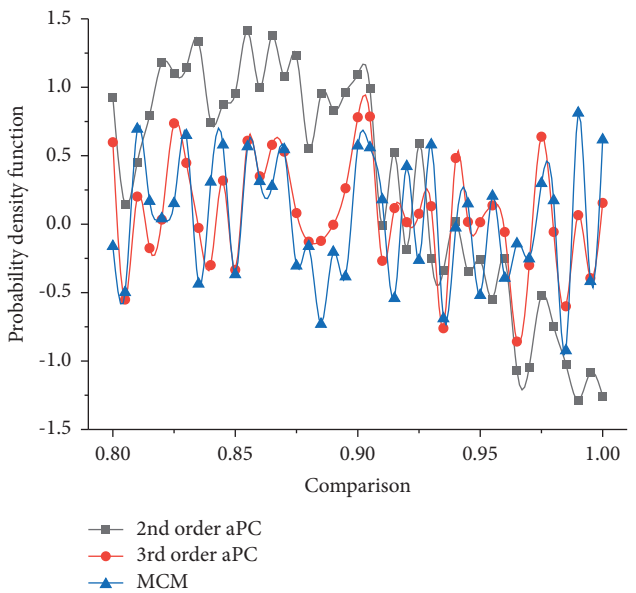

Figure 3: Comparison results of 2nd-order and 3rd-order APC.

within a fixed window size, which also causes the loss of semantic information to some extent because the order between words is not considered, and this part can compensate semantic information such as word order to some extent by combining lexical features. This part delves into the word vector learning model under the continuous bag-of-words model by combining the lexical feature information. According to the distributional hypothesis of words, words that are closer together have higher relevance, and there is also a certain distributional hypothesis between word properties according to the combination law of words [19]. The specific idea is that, based on the continuous bag-of-words model, the information of words, lexical properties, and their positions are fused, and the whole training corpus is preprocessed and labelled with lexical properties first; then, some basic work such as removing deactivated words is done, and, finally, the target words and their corresponding lexical properties are predicted by <word, lexical property> in the context, and this model framework is shown in Figure 4.

The model studies a training sample containing $2n+1$ words. These $2n+1$ words $(w_1, w_2, \ldots w_{2n+1})$, respectively, correspond to $\text{pos}_1, \text{pos}_2, \ldots \text{pos}_{2n+1}$, and the model in this part obtains the word vectors corresponding to the target words in the hidden layer by the positional weighting of the word vectors and the lexicality of the word vectors, and the main operation is in equation (3). To speed up the convergence of the model, this paper adopts the negative sampling technique strategy. The negative sampling technique is used to optimize the likelihood of positive and negative samples of a word and lexical information, so that the likelihood value of using positive samples is as high as possible, while the likelihood value of using negative samples is as low as possible. A positive sample can correspond to multiple negative samples, and this approach can effectively

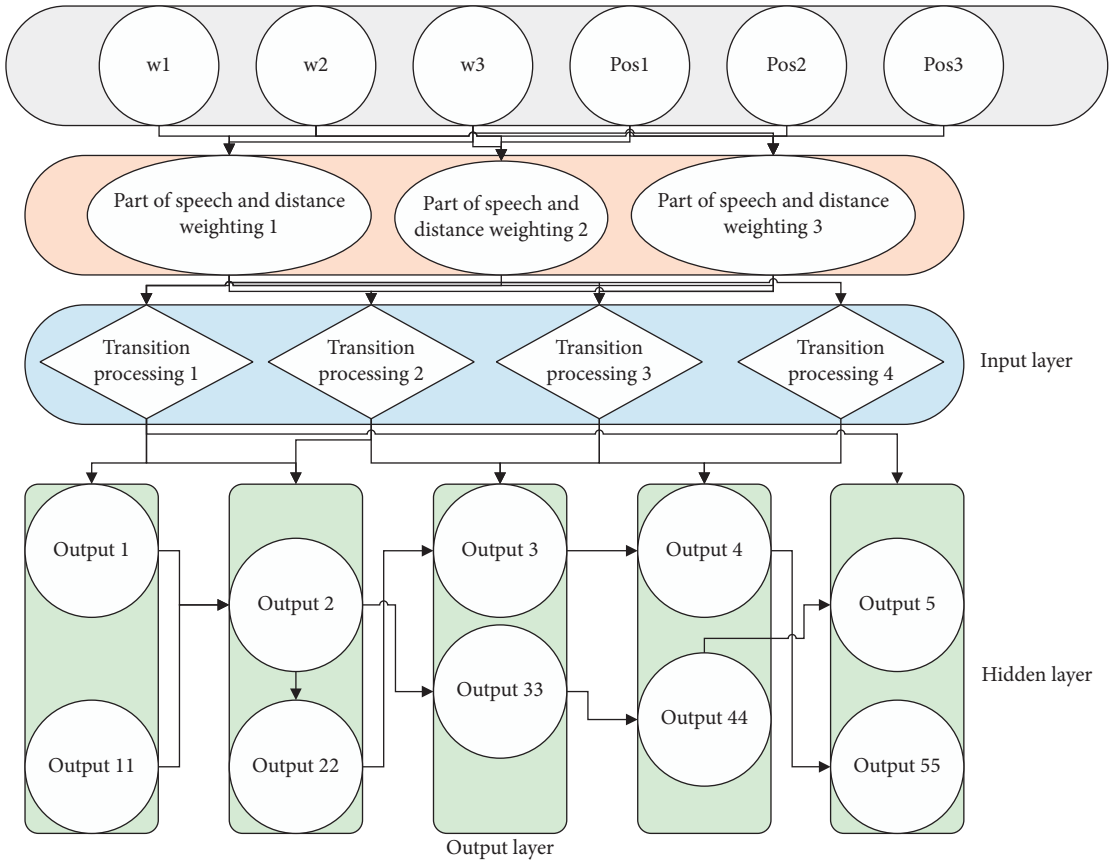

Figure 4: Structure of the English text summary extraction model.

improve the performance of the model to train word vectors on the whole training corpus.

$$x = \frac{t}{2t} \sum_{w_i \in a} p_i [e(w_i); e(\text{pos}_i)], \tag{3}$$

$$p_k(ab) = \frac{e_k(ab)}{n+k}. \tag{4}$$

By learning the word vectors in the previous part, a word vector obtained may have more than one-word vector, and if a dictionary of word vectors is constructed directly in this way, there will be a series of problems such as a large word vector table and a long query time. For this reason, in this part, by building a two-dimensional table representation of lexical and word vectors (as shown in Table 1), the time to find word vectors can be reduced, and thus the efficiency of processing can be improved. By creating a two-dimensional table, the word vectors can be quickly referenced by directly looking up the table during subsequent text processing.

This part deeply studies the word vector learning model under the continuous bag-of-words model by combining part-of-speech feature information. According to the distribution hypothesis of words, words with closer distances are more related, and, according to the combination rules of words, there is also a certain distribution hypothesis between parts of speech. This paper takes a single text as the research object and proposes an automatic extraction model of Chinese text summary based on fusion method based on corpus construction and preprocessing. Firstly, the text is analysed by dependent syntax according to the theory of dependent syntax analysis, and the text graph model is constructed by combining the text graph model with sentences as nodes; then the word cooccurrence network in sentences is constructed according to the cooccurrence theory in complex networks; next the topic model is constructed for the text, and, finally, the comprehensive scoring function of sentences is used to calculate the weights of sentences in the text and output the sentences within the text threshold as text summaries. It is proved that the method improves the readability of the text summary while providing enough information in the digest. Suppose that text T contains n sentences $\{s_1, s_2 \ldots s_n\}$ and m words. The traditional TF-IDF-based automatic text summary extraction method uses sentences as units, and the TF-IDF values corresponding to dissimilar words in the text obtained in Section 3.1 are used to calculate the sentence weights as shown in the following equation:

$$C_i = \sum_{j=0}^{i} T_j M_i. \tag{5}$$

In this paper, the local coherence model is applied to calculate the local coherence based on the relationship

Table 1: <word, lexical> word vector table.

Name	Part of speech 1	Part of speech 2	...	Part of speech n
<Word 1>	<Word 1, part of speech 1>	<Word 1, part of speech 2>	...	<Word 1, part of speech n>
<Word 2>	<Word 2, part of speech 1>	<Word 2, part of speech 2>	...	<Word 2, part of speech n>
<Word 3>	<Word 3, part of speech 1>	<Word 3, part of speech 2>	...	<Word 3, part of speech n>
<Word i>	<Word i, part of speech 1>	<Word i part of speech 2>	...	<Word i, part of speech n>
<Word n>	<Word n, part of speech 1>	<Word n, part of speech 1>	...	<Word n, part of speech n>

between adjacent sentences. Firstly, the syntactic components of the words in the sentences are obtained by using the dependent syntactic analysis, and a text graph model with sentences as nodes is constructed to calculate the local coherence of the sentences in the text; secondly, the complex network-based text word cooccurrence network model is used to obtain the comprehensive feature values of the network nodes in the text keyword extraction described in the previous paper to meet the demand for basic information of the text in the text summarization process; finally, a threshold value (limited to 10% in this paper) is set, and text sentences within the threshold value are selected as text summaries. The "Language Technology Platform (LTP)" is an open Chinese natural language processing system developed by the Social Computing and Information Retrieval Research Center of HUST. LTP develops an XML-based representation of language processing results and provides five core technologies for Chinese processing, including lexical, syntactic, and semantic. Kalman filter models are widely used to solve these dynamic state space models. It is a recursive method and there are different variants of these filter models due to the degree of linearity of the equations involved. In this paper, we call its syntactic analysis module to perform syntactic word separation on the text. From the analysis results, we can see that the core predicate of the sentence is "attach importance," the subject is "country," and the proposed object is "northeast China, etc." The object of "attach importance" is "work." Through the above syntactic analysis, it is easy to see that "attach importance" is "country," not "high" or "northeast." "Even though both "height" and "northeast" are nouns, they are even closer to "value." Natural language processing exists for machines that can better imitate human processing of natural language. The purpose is to complete tasks such as automatic speech dialogue and automatic text writing on large data and be as intelligent as the human brain.

$$\widehat{y} \le \delta\left(h_0 w_s^i + b_s\right). \tag{6}$$

4. Results and Analysis

4.1. Data-Driven Model Results. The data-driven polynomial chaos expansion method can be applied in the case of restricted statistics. It is necessary to point out that the main difference between the generalized polynomial chaos expansion method and the data-driven polynomial chaos expansion method is that the polynomial bases of the two are computed in different ways: the former is chosen from the Wiener-Askey mechanism, while the latter directly constructs the polynomial bases using several-order moment information of the statistics. To verify the validity of the road network travel time estimation model WCPR, an example study is conducted in this section using a large amount of sparse GPS trajectory data collected in the Beijing urban road network. The experimental data and its related preprocessing process are described in detail in Chapter 3 of this paper and will not be repeated here. In the following section, the experimental results of each component of the WCPR model, including travel time modelling, probabilistic traffic state clustering, and travel time estimation, are analysed and discussed in detail. To reduce the influence of abnormal data on the experimental results, two types of data are not considered in the experiments. One is the road segments that were passed less than 60 times (once a day on average) because these are the road segments that people rarely use and the data on these road segments are likely to be noisy data. The other is the drivers who do not generate trajectory data in all four time periods because these drivers are inactive in these periods, and modelling their travel time will not only increase the tensor sparsity but also have no practical meaning. After data cleaning, a total of 84,100 road segments and 29,083 drivers were used in the experiment. Therefore, the dimension size of the tensor is the product of 84100 and 29083. In addition, considering the error of data preprocessing, for each road segment, the travel time whose significance level is less than 0.05 is considered as noise and is screened out, as shown in Figure 5.

The experimental results of single-sentence summaries can be seen for the five baseline methods, CI, Text Rank, MMR, Lead Last, and Lead First, and the TSMMR model proposed in this paper, respectively. The results in the comparison show that the Lead Last method has the worst effect, which proves that most of the core statements of the articles are not placed at the end of the article in this dataset; among the other baseline methods, CI, which only considers the influence of keywords on the importance of sentences, has the worst effect, and the other three methods have better overall effect, but the difference is not large. The TSMMR model proposed in this paper has the best overall effect, with Rouge-L about 5 percentage points higher than the MMR algorithm, which has the best effect among the baseline methods, Rouge-2 about 3 percentage points higher than the Text Rank algorithm, which has the best effect among the baseline methods, and Rouge-1 about 5 percentage points higher than the MMR algorithm, which has the best effect among the baseline methods. This demonstrates that the TSMMR model proposed in this paper works best when only one sentence is extracted as a summary. Extremely long and extremely short travel time samples should be removed from

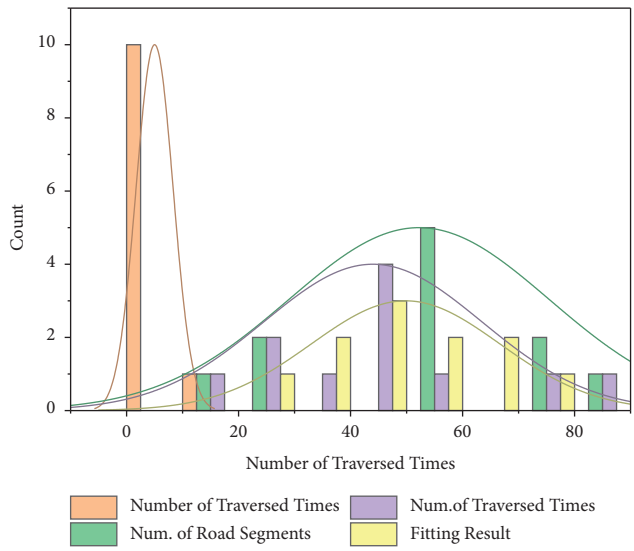

Figure 5: Trend of the number of test texts.

the database to obtain more reliable travel time estimates. However, identifying and dealing with invalid data from data are not always straightforward. Therefore, some methods have been proposed for the identification of invalid data, such as travel time data caused by halfway stops, duplicate data, data where the driving speed exceeds the speed limit value, and so forth.

In terms of the generalizability of the model, the TSMMR model proposed in this paper shows the best results in both single-sentence abstracts and multisentence abstracts experiments, indicating that the method can not only satisfy the task of extracting multisentence abstracts but also be extended to related tasks such as news headline generation. In addition, the model proposed in this paper is not limited to the domain of extracting article abstracts but can also be extended to the domain of extracting abstracts from other media. For example, in video subtitle summary extraction, the keywords contained in a single subtitle, the location information of the subtitle, and the similarity between the subtitle and all subtitles are used to extract the summary of video subtitles, so that users can understand the general meaning of the video in advance through the video subtitle summary and select more interesting contents; in video pop-up summary extraction, the keywords contained in the video pop-ups are counted by filtering the lexicality to extract the summary of the video pop-ups, so that users can understand each other. In the summary extraction of video pop-ups, the summary of video pop-ups is extracted by filtering the word nature and counting the keywords contained in the video pop-ups so that users can understand the evaluation of the video content by other users.

Autodesk's BIM series software, which is called Navisworks Manage software, contains progress simulation, collision detection, scene rendering, and other functions. The Clash Detective module in the software provides a very powerful collision detection function, which can check the design and construction errors in time and generate a collision check report with pictures and member coordinates information, including the content of model-to-model comparison collision, model internal member collision, gap distance collision between members, and so forth, combined with the structural model we created, using the software. The collision check is performed to verify the feasibility of our model and to check the engineering practicality of the data-driven approach. Another type of data mining model is based on artificial intelligence, mainly including artificial neural network, support vector machine, nonparametric regression model, wavelet analysis model, and deep learning model that has attracted more attention in recent years. When collisions of members are found, the model is modified to achieve the accurate model required for construction. Since Revit does not have construction rules that conform to our Chinese steel flat method, the calculated quantities are not following our national calculation rules when using the foreign Revit Structure structural software. In this paper, the data-driven model is based on the 11G101 plain-law drawing set, which is theoretically consistent with our domestic costing software, so the comparison of rebar calculation is carried out. Since there is only a breakdown of each bar in Revit, we need to sort out the breakdown of the bars, summarize the total amount of each type of bar, extract the number of bars of each length, and compare it with Quanta's calculation. From the figure, the calculated quantity of Revit is about 101.405 tons, and the quantity of Quanta's reinforcing steel is about 99.367 tons, with an error ratio of 2.05%, which is within 5%, and the calculation result is reliable. The results of its Revit rebar calculation are shown in Figure 6.

5. English Text Summary Extraction Model Application

We evaluate our approach on the English CNN/Daily Mail and DUC-2004 datasets, which are standard benchmark datasets for abstract text summarization. We choose CNN/Daily Mail as the training dataset, and the processing and detailed description can be found in the experimental section of Chapter 3. The average length of the abstracts is 52 words, and the average length of the original text is 685.2 words. The DUC-2004 dataset is an abstract evaluation set, which consists of 500 news articles [20]. Specifically, through the online and offline data of the controlled system, data-based establishment, adjustment, evaluation, optimization, and other functions are realized. When using the data-driven method for modelling, the data is used as the characteristic parameter of the object to control and generate the model, and the value of the characteristic parameter can also be reflected in the three-dimensional model. Through the results, we find that the advantage of this model is not obvious when the input source text is short, and the evaluation index value of this model is significantly higher when the length is more than 30 words, which indicates that this model is more accurate in acquiring the contextual semantics of long text, and the ability of long-range dependence is improved. When the input text is shorter, it is found that the evaluation index results values of the simple Transformer model and the PGEN model are similar, which

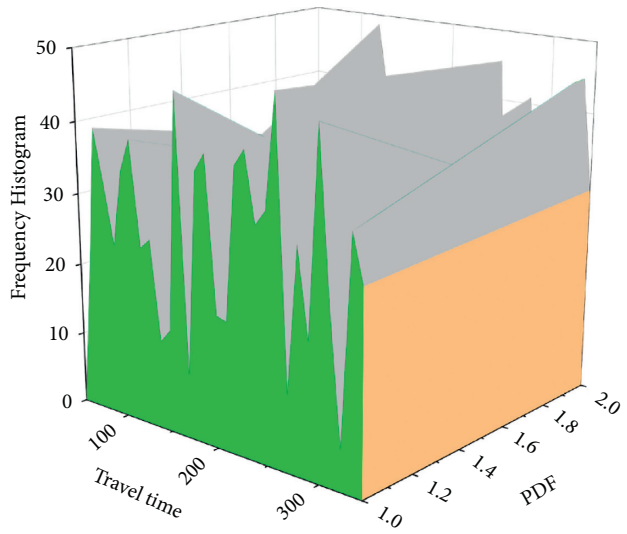

Figure 6: Data-driven distribution results.

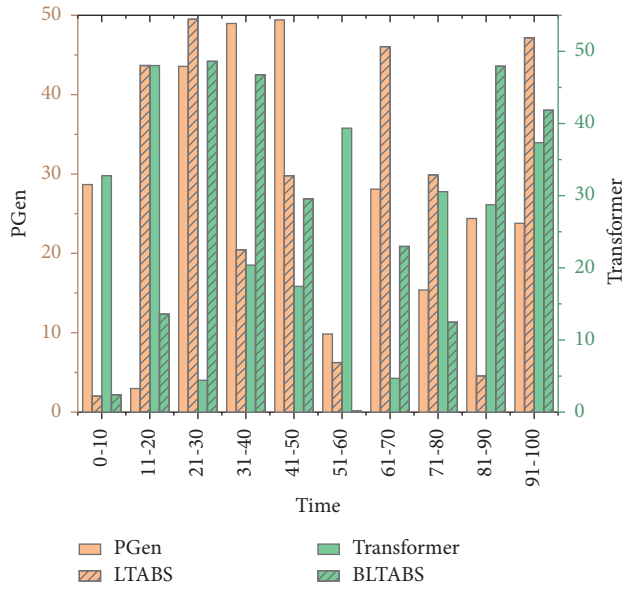

Figure 7: Input Rouge-1 index scores for different length sources.

indicates that the simple Transformer model is more capable of handling short texts and the generation capability can be comparable to the LSTM network with the addition of attention mechanism. After adding LSTM sequence information, there is a small improvement in the index value, which indicates that sequence information has a necessary role for text data. By adding different factors to test in texts of different lengths, it is proved that the model in this paper has some advantages for handling longer texts, and, by improving the Transformer model, it can be more suitable for the summary generation task, as shown in Figure 7.

Word frequency refers to the number of times a word appears in a document, and it is usually assumed that the more often a word appears in each document, the more likely that word is the core word of that document. The calculation of word frequency statistics is relatively simple, but if the articles in a corpus are discussing a common topic, only with a different focus, then words related to this topic will appear in high frequency, and, for a particular article, these words are not necessarily the core words of that article. So, the simple extraction of keywords based on word frequency statistics may not be very accurate. TFIDF, on the other hand, improves the shortcomings of word frequency statistics by calculating the inverse document frequency of words in addition to considering word frequency. The basic idea is that if a word appears in most of the articles in the corpus, its TFIDF value is not necessarily high even if the word has a high frequency. The method of extracting keywords by TFIDF is still used in a variety of extractive text summarization algorithms. Topic modelling is one of the powerful tools for text mining, which can mine potential connections between data and between data and text through prior knowledge of text. Especially when dealing with source text of discrete data, topic model can have its biggest advantage.

The core idea of the keyword extraction method based on the graph model is to construct a word network graph with words as nodes and relationships between words as edges and then select words with important roles on the network graph as keywords by analysing the graph. The words in this network are usually preprocessed and filtered words, such as removing deactivated words, removing low-frequency words, or removing certain lexical words. The word nodes usually construct edges between them by their cooccurrence or syntactic relationships. The text summarization task aims to convert text into short summaries containing key information. Today's automatic text summarization methods are mainly divided into extractive models and generative models. Cooccurrence means that, by setting a window of length K if two words appear in this window, there is a cooccurrence relationship between these two words, and the weight of the edge is determined by the distance between the two words or the number of appearances in multiple windows; syntactic relationship means whether the structure between two words is subject-predicate, verb-object, and so forth. Word networks generally consist of nodes, edges, and the weights of edges. The weights of edges are generally called degrees and are classified into directed and undirected graphs according to whether they are directed or not. The importance of a node consists of various features; its features, such as word frequency, word nature, and position of words, are called local features; features influenced by other words (or nodes) are called association features. The final important nodes, that is, keywords, are extracted by calculating the total scores of the nodes' local features and the association features obtained from the association of other words in the graph model, and the results of the application of the data-driven English text summary extraction model based on the data are shown in Figure 8.

The keyword features of the Chinese and English corpus were extracted separately, and the results of mutual transformation by bilingual word vector alignment (BWR) and machine translation (MT), where Fn refers to the keywords of the English corpus, Fen-cn represents the keywords of the

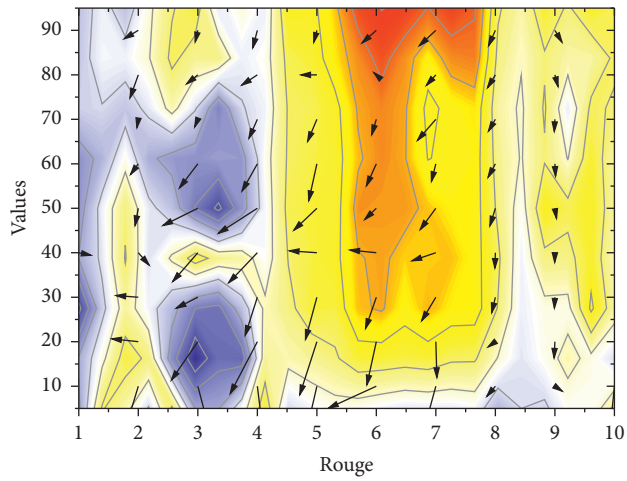

Figure 8: Application study results.

English corpus transformed into Chinese by two ways, Fcn refers to the keywords of the Chinese corpus, and Fcn-en represents the keywords of the Chinese corpus transformed into English by two ways. Computers cannot accurately understand the meaning and generate their own cognition after reading some related texts like humans. They can only use statistics, machine learning, simple reasoning engines, primary memory mechanisms, and other methods to correlate documents. Processing can generally only be extracted from documents or simply "thinking processing" to form the final abstract of the article. By manually comparing the results of machine translation and bilingual word vector alignment for each word pair, most of them are correct, and those marked in bold in the table are the word pairs with problems. In the word pairs translated from Chinese to English, there are also some errors in the translation of "year, month, and day," both of which are not well translated but correspond to some specific dates; in the translation of the word "China," the machine translation is more accurate. "China" is written with "C" capitalized, while the bilingual word vector alignment is transformed into "china" ("c" lowercase) which deviates from the original meaning; in addition, for most of the verbs, both ways can only be transformed into one morphology, and most of them are in the past tense, which is in line with the morphology of most of the verbs in the news data of this paper.

In this part, we propose a two-stage summary generation method that combines the advantages of two summary methods, and the specific idea is that the first stage extracts some important sentences of the document by combining "pseudoheadings," which fully considers the sentence position, paragraph in the second stage; the extracted sentences are reorganized and rewritten using the beam search algorithm to generate new sentences as "pseudotitles" for the next stage of the document. The first and second steps are cyclically executed until the optimal "pseudotitle" is obtained for the whole document, and the "pseudotitle" that satisfies the final condition is used as the final summary of the document. Extensive experiments are conducted on the corresponding English and Chinese datasets, and the experiments show that the method can obtain better summary results.

6. Conclusion

In this paper, we propose a data-driven modelling approach based on the theory of data-driven uncertainty analysis by parametrically designing the model components and then performing data-driven analysis on them and finally using Revit as a carrier for secondary development of the parametric components to achieve data-driven modelling. To obtain a higher and more suitable word vector representation for summarization, the paper proposes a fine-grained word vector representation method incorporating lexicality, as representation learning is a fundamental task for conducting natural language processing and a cornerstone for conducting subsequent research on natural language-related tasks. By combining lexical and location information, this paper constructs a new, fine-grained word vector representation for text summarization and combines the two-dimensional table representation of <word, lexical> word vectors to reduce the size of the word vector lookup table and improve the query efficiency, and experiments show that the proposed method has better semantic representation of text. Since the existing methods mostly focus on the amount of text information contained in the digest and ignore the coherence of the digest itself, this paper combines the text graph model, complex network theory, and LDA topic model to construct a sentence synthesis scoring function to calculate the text single-sentence weights and output the sentences within the text threshold in a descending order as the text digest. The algorithm improves the readability of the digest while providing enough information in the digest. In the next study, the semantic analysis of the text will be enhanced to further improve the semantic information of the digest; in addition, the self-built corpus can be expanded to explore the improvement of the accuracy and readability of this paper's method for other types of Chinese text digests.

Data Availability

The data used to support the findings of this study are available from the author upon request.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This work was supported by Wuhan Huaxia University of Technology.

References

[1] C. Guan, J. Mou, and Z. Jiang, "Artificial intelligence innovation in education: a twenty-year data-driven historical analysis," *International Journal of Innovation Studies*, vol. 4, no. 4, pp. 134–147, 2020.

[2] A. Akbari, "Translation quality research: a data-driven collection of peer-reviewed journal articles during 2000–2017," *Babel. Revue internationale de la traduction/International Journal of Translation*, vol. 64, no. 4, pp. 548–578, 2018.

[3] G. M. D. S. Ferreira, L. A. D. S. Rosado, M. S. Lemgruber, and J. D. S. Carvalho, "Metaphors we're colonised by? The case of data-driven educational technologies in Brazil," *Learning, Media and Technology*, vol. 45, no. 1, pp. 46–60, 2020.

[4] P. Bauer and G. Anzer, "Data-driven detection of counterpressing in professional football," *Data Mining and Knowledge Discovery*, vol. 35, no. 5, pp. 2009–2049, 2021.

[5] R. Tachicart and K. Bouzoubaa, "Moroccan data-driven spelling normalization using character neural embedding," *Vietnam Journal of Computer Science*, vol. 8, no. 01, pp. 113–131, 2021.

[6] A. Piotrkowicz, K. Wang, J. Hallam, and V. Dimitrova, "Data-driven exploration of engagement with workplace-based assessment in the clinical skills domain," *International Journal of Artificial Intelligence in Education*, vol. 31, no. 4, pp. 1022–1052, 2021.

[7] A. M. Shahat Osman and A. Elragal, "Smart cities and big data analytics: a data-driven decision-making use case," *Smart Cities*, vol. 4, no. 1, pp. 286–313, 2021.

[8] S. Heldens, P. Hijma, B. V. Werkhoven, J. Maassen, A. S. Z. Belloum, and R. V. V. Nieuwpoort, "The landscape of exascale research: a data-driven literature analysis," *ACM Computing Surveys (CSUR)*, vol. 53, no. 2, pp. 1–43, 2020.

[9] Z. Xu and Y. Dang, "Automated digital cause-and-effect diagrams to assist causal analysis in problem-solving: a data-driven approach," *International Journal of Production Research*, vol. 58, no. 17, pp. 5359–5379, 2020.

[10] Z. Xu, Y. Dang, P. Munro, and Y. Wang, "A data-driven approach for constructing the component-failure mode matrix for FMEA," *Journal of Intelligent Manufacturing*, vol. 31, no. 1, pp. 249–265, 2020.

[11] C. Lim and P. P. Maglio, "Data-driven understanding of smart service systems through text mining," *Service Science*, vol. 10, no. 2, pp. 154–180, 2018.

[12] S. Zhang, E. J. M. Carranza, H. Wei et al., "Data-driven mineral prospectivity mapping by joint application of unsupervised convolutional auto-encoder network and supervised convolutional neural network," *Natural Resources Research*, vol. 30, no. 2, pp. 1011–1031, 2021.

[13] N. Sun, J. Zhang, P. Rimba, G. Shang, Y. Z. Leo, and X. Yang, "Data-driven cybersecurity incident prediction: a survey," *IEEE communications surveys & tutorials*, vol. 21, no. 2, pp. 1744–1772, 2018.

[14] D. Spry, "Facebook diplomacy: a data-driven, user-focused approach to Facebook use by diplomatic missions," *Media International Australia*, vol. 168, no. 1, pp. 62–80, 2018.

[15] X. Hu, L. Ding, J. Shang et al., "Data-driven approach to learning salience models of indoor landmarks by using genetic programming," *International Journal of Digital Earth*, vol. 13, no. 11, pp. 1230–1257, 2020.

[16] J. M. Cole, "A design-to-device pipeline for data-driven materials discovery," *Accounts of Chemical Research*, vol. 53, no. 3, pp. 599–610, 2020.

[17] W. Yu, M. A. Jacobs, R. Chavez, and M. Feng, "Data-driven supply chain orientation and financial performance: the moderating effect of innovation-focused complementary assets," *British Journal of Management*, vol. 30, no. 2, pp. 299–314, 2019.

[18] B. Zhao, F. Yang, R. Zhang, J. Shen, J. Pilz, and D. Zhang, "Application of unsupervised learning of finite mixture models in ASTER VNIR data-driven land use classification," *Journal of Spatial Science*, vol. 66, no. 1, pp. 89–112, 2021.

[19] W. L. Johnson, "Data-driven development and evaluation of Enskill English," *International Journal of Artificial Intelligence in Education*, vol. 29, no. 3, pp. 425–457, 2019.

[20] D. Hooshyar, M. Yousefi, and H. Lim, "A systematic review of data-driven approaches in player modeling of educational games," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 1997–2017, 2019.