ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 72
IV International Research Conference "Information Technologies in Science, Management, Social Sphere and Medicine" (ITSMSSM 2017)

Text structures synthesis on the basis of their system-forming characteristics

Liubov. S. Lomakina
Institute of Radio Electronics and Information Technologies
Nizhny Novgorod State Technical University n.a. R.E. Alekseev, NNSTU
Nizhny Novgorod, Russia
llomakina@list.ru

Anna. S. Surkova
Institute of Radio Electronics and Information Technologies
Nizhny Novgorod State Technical University n.a. R.E. Alekseev, NNSTU
Nizhny Novgorod, Russia
ansurkova@yandex.ru

Dmitry. V. Zhevnerchuk
Institute of Radio Electronics and Information Technologies
Nizhny Novgorod State Technical University n.a. R.E. Alekseev, NNSTU
Nizhny Novgorod, Russia
zhevnerchuk@yandex.ru

Igor. D. Chernobaev
Institute of Radio Electronics and Information Technologies
Nizhny Novgorod State Technical University n.a. R.E. Alekseev, NNSTU
Nizhny Novgorod, Russia
ichernobnn@gmail.com

***Abstract*—This paper deals with different text structure synthesis methods using the concept of hidden parameters. Three main methods of text structure synthesis using the concept of hidden parameters were described in this paper: information, parametric and non-parametric. The hidden parameters concept implies the texts structural invariants identification, which allows considering only meaningful characteristics for solving a particular problem. The concept of hidden parameters allows text structures patterns detection in the form of invariants and forming summarizing text model as a multidimensional object.**

Keywords— text analysis, text data, parametric synthesis, non-parametric synthesis, information synthesis

I. INTRODUCTION

Nowadays need in effective text information processing methods is increasing due to enormous amounts of text data in the digital form. Tasks of text structures analysis and synthesis, that is, a set of stable links of the text description features that provide its integrity and basic features preservation, are very important.

The article deals with features of the information, parametric (classical) and nonparametric synthesis of text structures from the point of the hidden parameters concept. Hidden parameters are the system-organizing characteristics of the object and characterize text structures. The hidden parameters concept allows discovering the text structures patterns in the form of invariants (universal, author's, topic) and forming summarizing text model as a multidimensional object, investigate its functioning mechanisms in the tasks of clustering, classification and identification of texts [1]. Classification refers to texts assignment to predefined classes, clustering - division of the given texts set into homogeneous by some criteria groups and identification refers to revealing the identity of objects or their attributes or properties.

Parametric and nonparametric synthesis are traditionally used in the multidimensional data processing tasks. Parametric synthesis can be described as the process of determining parameters of elements of a synthesized object with a predetermined object structure. Nonparametric synthesis – as the process of determining the structure of an object and the parameters values of its constituent elements. Information synthesis plays especial role in text data processing. The information measure is used in the synthesis to determine the structure, parameters and other characteristics of text elements as a criterion.

II. PARAMETRIC SYNTHESIS

The parametric synthesis methods are based on the assumption that text data described by distributions from parametric families, that is, under the given distribution law assumption. Therefore the task is in determining of unknown distribution parameters. Multidimensional normal distributions are considered as usual. Often the hypothesis that covariance matrices for different classes coincide is accepted.

The distribution densities or their estimates usage is based on the Neumann-Pearson lemma, according to which the most powerful criterion for testing simple statistical hypotheses is the likelihood ratio [2].

Suppose that in some n-dimensional features it is possible to characterize a set of objects by unknown probability distribution $P(x)$. For training set X the appearance of elements of which also corresponds $P(x)$, the distribution by given k classes is known. That is, there is some variable y that determines the assignment of an object to a class x:

$$y_s(x) = \begin{cases} 1, \text{if } x \text{ belongs to class } s; \\ 0, otherwise, s = \overline{1,k}; \end{cases} \quad (1)$$

From the point of the hidden parameters concept, the variable y reflects the hidden parameter – the invariant of the class, corresponding to the elements from one class.

For clarity, we consider the case of two classes, which can be extended for the case with many classes. Suppose there are defined two sets of objects representing a points in n-dimensional space W and the probability densities $f(x)$ and $g(x)$ correspond to these sets are known [3]. It is necessary to build a separation in space in that way so as many points from one class lying on the one side of the boundary and from other class – on the opposite side. That is, the boundary defines an area such that:

$$\int_R g(x)dx = \int_{W-R} f(x)dx = 1 - \int_R f(x)dx \quad (2)$$

$$\text{i.e. } \int_R (f(x) + g(x))dx = 1 \quad (3)$$

To build a separation, it is necessary to minimize the probability of false classification:

$$\int_R g(x)dx \to \min \quad (4)$$

On the basis of equation (3) minimization criterion can be written:

$$\int_R (\beta g(x) - f(x))dx \to \min \quad (5)$$

The desired partition can be obtained if to the set R there will be assigned points for which the condition is true:

$$\beta f(x) - g(x) < 0 \quad (6)$$

In that case the boundary R is given by equation, i.e. it is determined by the likelihood ratio:

$$\beta = \frac{f(x)}{g(x)} \quad (7)$$

The parametrical synthesis of text structures includes most methods and algorithms for data classification such as the supervised learning. It is assumed that the general structure and classification criteria are known in advance and during the synthesis process specific parameters of the objects of each class are determined and according to the revealed data, a classification of unknown texts is performed. The parameters of objects and the structure of classes depend on classification tasks and classification criteria: thematic, linguistic, stylistic, etc.

The methods of parametric synthesis can also be attributed with a classification using recurrent neural networks (RNN). Neural networks with recurrent architecture are successfully used to solve problems in machine learning. They are so called due to the presence of feedback in the neuron, which allows taking into account its previous states.

The hidden state of a recurrent neuron contains the information about its previous hidden state and also known as the RNN's "memory". The memory allows taking into account previous elements from training data sequence.

Such networks outperform traditional feedforward nets and gain the ability to get the better models. Nevertheless, there are some constraints imposed on the processed data sequence length. These constraints are linked with the common for RNN and feedforward neural networks issue of vanishing gradient. This issue results in the freezing of the network learning process.

There are some key points in the network learning process:

1. Vectors of the weight coefficients initial initialization. It is recommended [4] to use a sample of random variables in range $[-\frac{1}{\sqrt{n}}; \frac{1}{\sqrt{n}}]$, where n – vector dimension.
2. Applying a cost function to calculate net error. The cross entropy function can be used as a cost function: $L(y,o) = -\frac{1}{N}\sum_{n \in N} y_n \log o_n$ (8), where y – correct sequence, o – the network output, n – the index at a sequence, N – the number of sequences.
3. Gradient calculation of the network hidden parameters.
4. Applying gradients for hidden parameters adjustment.

The aim of network learning process is to determine the vectors of weight coefficients for the loss function

minimization by any optimization algorithm, for example, by the stochastic gradient descent method (SGD). During learning process, training data is transferred in the loop to the network input. One complete cycle iteration is called an epoch. At each epoch the back propagation through time (BPTT) algorithm is performed, in which the gradients of the vectors of the weight coefficients are calculated. After BPTT completion, vectors of the weight coefficients are adjusted up by the following equation:

$$\begin{cases} W -= \nu * \frac{dL}{dW} \end{cases} \quad (9)$$

where $\frac{dL}{dW}$ – the obtained gradient, ν - the learning step.

The Long-Short-Term-Memory (LSTM) network [5] is quite popular kind of the RNN. In addition to the recurrent connection and, as a consequence, its memory, network is designed to reduce the impact of the vanishing gradient problem. This is achieved through the filters usage.

The LSTM cell structure is shown on Fig. 1.

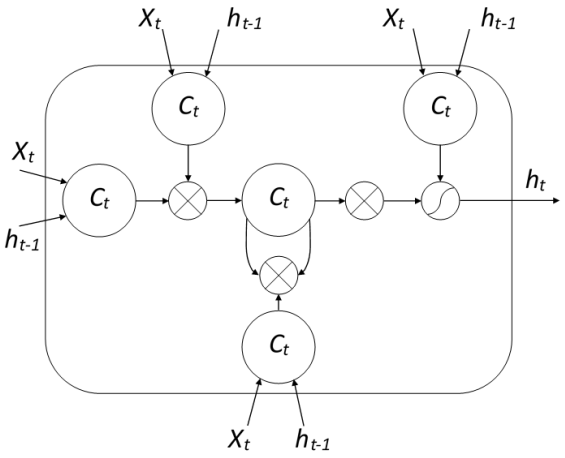

Fig. 1. The LSTM cell structure.

The LSTM equations are:

$$i_t = \sigma(U_i x_t + W_i h_{t-1}) \quad (11)$$

Here σ – the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ (12)

$$f_t = \sigma(U_f x_t + W_f h_{t-1}) \quad (13)$$

$$o_t = \sigma(U_o x_t + W_o h_{t-1}) \quad (14)$$

$$g_t = tahn(U_g x_t + W_g h_{t-1}) \quad (15)$$

$$c_t = i_t \, o \, g_t + f_t \, o h_{t-1} \quad (16)$$

$$h_t = o_t \, o tahn(c_t) \quad (17)$$

Where h_t – the cell output, tanh is the hyperbolic tangent function:

$$\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1} \quad (18)$$

and o is the element wise product operation, c_t - the internal state of the cell.

Gates i, f, o are input, forget, output gates respectively. The sigmoid function squashes values of passed vectors to the interval from 0 to 1. Threshold for passing a vector through a gateway is determined by gates and other vectors elementwise multiplication. Thus, gates regulate how much of vectors are passed through them. A value of 0 closes the gate and a value of 1 opens it. The vector g contains new candidate values scaled by the forget gate. The cell state c – its internal memory. It is the product of the previous state and the forget fate. Based on its memory the cell calculates its output.

In addition to neural networks, parametric methods can also be attributed with Support Vector Machines (SVM), Mountain Clustering method, fuzzy c-means (FCM) method, Kernel Fuzzy Clustering method, the K-nearest neighbors (kNN) method, decision trees based methods.

III. NONPARAMETRIC SYNTHESIS

In non-parametric synthesis, to restore an unknown structure, a set of objects with the same structure is used as a training sample (by some characteristics). In the training process, the parameters common to all objects, which make it possible to synthesize the structure necessary to solve the task, are singled out. These parameters allow to carry out the synthesis of the structure which is necessary to solve the task.

In case of non-parametric synthesis, regardless of the source task it is necessary to perform a preliminary data analysis and carry out preliminary data structuring with minimal a priori information about the investigated objects. Most often different clustering algorithms are used for this purpose. In this case, when specifying a particular clustering method, specifying the metric, the parameters of the algorithm, a choice of a generalized structure of object occurred. When evaluating clustering methods, it is required to estimate the "naturalness" of the resulting objects partition [6]. In this case, for different source data, different algorithms can best perform clustering.

When considering such complex system objects as text, it is possible to divide texts according to various features, i. e. creating several natural classifications that can be obtained by revealing the hidden parameter corresponding to the problem being solved. This revealing is performed by hidden parameter value definition for considered texts and by division by groups, depending on the value of this parameter. In this case, it is possible a specific situation in which a specific hidden parameter is not explicitly determined, but reflected through other observable parameters. With this, depending on the used models, different clustering algorithms can give different portioning results, because these models reflect various hidden parameters in different degrees.

One of the universal approaches to solving non-parametric synthesis tasks is the Expectation-Maximization (EM) algorithm usage [7]. It is based on the maximum likelihood criteria for unknown parameters Θ. Here observable parameters X, unobservable hidden parameters Z and the likelihood function

$$L(\Theta, X) = p(X \mid \Theta) = \sum_{Z} p(X, Z \mid \Theta) \quad (19)$$

are considered.

Each iteration of an EM algorithm consists of two steps. At the Expectation step the expected value is calculated in the hidden variables vector Z by the current approximation of the parameter vector Θ, which corresponds to the assignment of object of classes. At the Maximization step the maximum likelihood estimation is evaluated, its maximum is determined and the model parameters are recalculated: the following parameter value Θ^{i+1} is found by the current values of Θ^{i} and Z.

Besides EM method neural networks with recurrent architecture are also capable to solve a non-parametric synthesis problem. The main equations for a simple RNN are:

$$s_t = \tanh(U * x_t - W * s_{t-1}) \quad (20)$$

$$o_t = soft\max(V * s_t) \quad (21)$$

where softmax is the normalized exponential function

$$soft\max(z) = \frac{e^{z}t}{\sum_{n=1}^{N} e^{z}n} \quad (22)$$

$$z = V * s_t \quad (23)$$

Here s_t is the recurrent neuron hidden state, x_t - net input at step t. Net output at step t o_t is calculated based on its hidden state. Parameters U, V, W are the vectors of the weight coefficients. In contrast to traditional networks that use different parameters at each layer, RNN shares same parameters U, V, W across all steps. This allows to reduce common quantity of parameters during learning process.

After BPTT completion each network parameter is adjusted up by the following equations:

$$o\begin{cases} U -= \nu * \frac{dL}{dU} \\ V -= \nu * \frac{dL}{dV} \\ W -= \nu * \frac{dL}{dW} \end{cases} \quad (24)$$

where $\frac{dL}{dU}$, $\frac{dL}{dV}$, $\frac{dL}{dW}$ – calculated gradients, ν - the learning rate.

Most algorithms for nonparametric text structures synthesis are relational methods, based on the mutual texts relation and difference between objects determining. The usage of known nonparametric methods such as the method of signs or ranks, as well as methods based on the Kolmogorov complexity concept usage, makes it possible to estimate the unknown hidden structure of the of objects interconnection. Nonparametric methods include clustering methods based on neural networks (GSOM), methods of clustering based on fuzzy relationships and many others. An excellent style manual for science writers is [7].

IV. INFORMATION SYNTHESIS

A special place is occupied by information synthesis, which is not a contradistinction to parametric and nonparametric synthesis. It uses information metrics (information measure), which are convenient for application from the point of view of the interpretation of the physical meaning of phenomena. Using the information measure, it is determined what is common between the structures of objects in terms of the amount of information. The information approach usage makes it possible to simplify complex estimates of probability, in particular, to estimate the uncertainty by one value.

Algorithms of parametric and nonparametric synthesis can apply information, i.e. to refer to information synthesis. The distribution law determines indeed all information about the object. However, often the use of the amount of information is more convenient in practice. It can be said that algorithms of information synthesis make it possible to make use of all available information for the solution of tasks most fully. For the considered text data, the information component is of great importance, so the information synthesis algorithms were separated into a separate group for separate consideration. In this case, methods based on the information usage can be applied to solve any problems (classification and clustering). However, the information measure makes it possible to construct effective algorithms for solving identification problems.

Suppose x_i – an element of the text, N - the number of different values that an element x_i can take, $p(x_i)$ - the unconditional probability of an element x_i appearance in the text, and $p(x_i x_j)$ – is the probability of the pair of elements x_i and x_j appearance. Then it is possible to put a quantitative measure of the mutual information between elements in each pair. This measure is calculated by the formula:

$$I(x_i, x_j) = \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)}, i, j = \overline{1, N} \quad (25)$$

Information synthesis can be considered as a generalizing procedure of parametric and nonparametric methods and is applied to solve identification problems. Indeed, if the identification task is defined as the task of determining meaningful parameters for the source data (texts) and determining the specific values of the selected features, then

the task of the identification procedure has the features of both clustering (characterization) and classification (calculation of characteristic values). Parametric methods of identification include of text analysis methods based on the information approach, identification methods based on neural networks, text analysis methods based on entropy characteristics.

V. DISCUSSION AND EVALUATION

Classification of documents by thematic categories was carried out according to the methodology [8] using the proposed texts model in the form of N-gram spectra instead of the generally accepted word vectors or phrases. The check carried out on a specific body of texts showed an increase in classification efficiency due to the use of a more accurate text model instead of using resource-intensive classifiers or increasing the volume of training sets of documents. The quality of classification by different methods for the proposed model is presented in Table 1 (the number of correct classifier decisions was estimated).

With use of the LSTM network, a film reviews classification was performed. Each review is marked by sentiment (positive, negative) and presented as a sequence of word indices in dictionary of the entire dataset.

TABLE I. CLASSIFICATION RESULTS

Classifier	Quality of classification, percent
The Method of Support Vector Machine	100
The probabilistic neural network	99
Multilayer neural network	98
Decision rules	94
Decision Trees	94
LSTM	88
Naive Bayesian Classifier	86

Texts modeling methods based on entropy characteristics were used to classify texts of patents and traditional (artistic) texts [9]. The entropy characteristics were calculated for the level of symbols (letters) and the level of words. As shown by experiments, the characteristics at the level of letters do not allow to classify texts with the necessary accuracy: the quality of classification of texts is below 40%. The quality of the artistic texts classification using entropy characteristics calculated at the level of words amounted to about 80% of the correct decisions of the classifier, and when classifying patent texts – about 25%. Thus, the entropy characteristics at the word level reflect the author's style, being the author's invariants. At the same time, the texts of patent documents are subject to certain, pre-determined requirements and possess less vocabulary diversity, unlike artistic texts.

Nonparametric synthesis methods were used in the classification of users belonging to the social community, according to their text messages [10]. It was shown that when working with Internet texts, one should use methods of fuzzy clustering and use such parameters characterizing the authors of text messages as authoritativeness and number of followers.

TABLE II. CLUSTERIZATION RESULTS

Clusteringresults	Compression quantity, percent
PPMd	84
PPMd; Fast Clustering	84,5
LZMA	85,5
LZMA; Fast Clustering	85
BWT on Huffman algorithm	82,5
BWT on Huffman algorithm; Fasr Clustering	83,5

Based on the Kolmogorov's concept, a hierarchical text clustering method was proposed [11]. Table 2 represents the results of experiments on data clustering, depending on the compression algorithm used using the diagonal distance matrix (Fast Clustering).

The model of a set of texts was trained based on a recurrent neural network. It was shown that as the number of epoch's increases, the network error decreases and the generated text quality improves. The RNN learns better the smaller volume of texts and for model comparison, it is necessary to take models trained on the same data.

The information measure was used to perform the author of the source code texts identification based on the neural network technologies and entropy characteristics [12, 13]. The author of artistic texts identification based on the information portraits of texts comparison [14]. In determining the authorship of artistic texts, mutual information was calculated by such characteristics as individual letters in a raw, letters through one symbol, bigrams (two-letter combinations), and vowel letters. In order to characterize the texts proximity, the correlation coefficient K and standard deviation σ^2 were used. Identification results are given in Table 3.

TABLE III. IDENTIFICATION RESULTS

Identification results	Quality of identification, percent
Letters in a raw	75
Letters through one symbol	75
Bigrams	53
Vowel letters	82,2

VI. CONCLUSION

The paper considers information, parametric and nonparametric synthesis of text structures from the perspective of the hidden parameters concept. The hidden parameters concept implies the identification of texts structural invariants, which allows considering only the meaningful characteristics for solving a particular problem, thereby reducing the overall

dimension of the problem. It seems promising to create an open information system for text analysis, which makes it possible to implement a set of methods for synthesizing text structures for solving problems of processing texts of different types (artistic, scientific and Internet texts, texts of programs). An open information system can include specialized services for analysis and processing of texts, which allows real-time text data processing.

REFERENCES

[1] D. V. Lomakin, M. D. Lomakina, A. S. Surkova, "Forming methodology of system-organizing text data characteristics," Fundamental research, vol. 3, no. 11, pp. 480-83, 2015.

[2] E. I. Lehman, J. P. Romano, Testing statistical hypotheses. Springer Science & Business Media, 3rd ed. Mart 2006, pp. 786.

[3] M. G. Kendall, A. Stuart, The Advanced Theory of Statistics. Design and Analysis and Time-Series. Moscow: Science, pp. 585, 1976.

[1] X. Glorot, Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, vol. 9, pp. 249-256, May 2010.

[2] S. Hochreiter, J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, November 1997.

[3] A. I. Orlov, Applied Statistics – The State and the properties. Moscow: Exam, pp. 656, 2004.

[4] S. Russell, P. Norvig, Artificial Intelligence: A modern approach, 2nd ed. Pearson Education, 2003.

[5] L. S. Lomakina, A. V. Mordvinov, A. S. Surkova, "Construction and investigation of the model text for its classification by subject categories," Control Systems and Information Technologies, vol. 43, no. 1, pp. 16-20, 2011.

[6] L. S. Lomakina, S. S. Surkova, "Applied aspects of conceptual analysis and modeling of text structures," Fundamental research, vol. 7, no. 3, pp. 540-544, 2015.

[7] L. S. Lomakina, A. S. Surkova, S. S. Budenkov, "Clustering text data based on fuzzy logic," Control Systems and Information Technologies, vol. 55, no. 1, pp. 73-78, 2014.

[8] L. S. Lomakina, V. B. Rodionov, A. S. Surkova, "Hierarchical clustering of text documents," Automation and Remote Control, vol. 75, no. 7. pp. 1309-1315, 2014.

[9] A. S. Surkova, A. A. Tsarev, "Application of neural networks for source code authorship identification," Control Systems and Information Technologies, vol. 63, no. 1, pp. 78-82, 2016.

[10] M. S. Sementsov, A. S. Surkova, "Entropic characteristics of symbolic diversity in the texts of the program source codes," Control Systems and Information Technologies, vol. 59, no. 1.1, pp. 173-176, 2015.

[11] A. S. Surkova, "Text authorship attribution on the basis of information portraits," Vestnik of Lobachevsky University of Nizhni Novgorod, no. 3-1, pp. 145–149, 2014.