Model of Forecasting the Social News Events on the Basis of Stochastic Dynamics Methods

D.O. Zhukov [1]*, *A.M.* Zamyshlyaev[2], and *O.A.* Novikova[1]

[1] Moscow Technological University, Moscow, Russia
[2] Researches and Design Institute of Information Technology, Automation and Telecommunications on Railway Transport (JSC "NIIAS"), Moscow, Russia

Abstract. For a description of the information space it is introduced a vector representation of the constituent text documents that are bound by the events described in the timeline. The predicted event is also represented by a vector obtained on the base of its text description. The mean value of projections of the information space in the direction of the vector of predicted events at different time points is considered as a set of information system states. It is also entered the change values of states. To describe transitions between states is used a probabilistic approach and the difference transition scheme. This makes it possible to get the dependence of the time for the value of the probability density for the event "detection information system in a state" in the form of a second order differential equation. On the basis of this equation is formulated and solved the boundary problem. Carried out by the authors the analysis of the stochastic dynamics of achievement a threshold of realization of news events has allowed the establishing of the ability to increase the probability of transition almost simultaneously with the beginning of the process of the news cluster structure changing. This is due to the presence of the memory of previous states in the information system and the possibility of self-description, as a result of accounting in the differential model information processes on the basis of the second derivative over time. In addition, the proposed model demonstrates the possibility of sudden changes in the probability of crossing the threshold of events and takes into account the presence of oscillations in her behavior. Based on the model developed it is proposed the algorithm for analysis of news clusters relationship in the information field with the possibility of occurrence of the predicted event, and determined the possible time of its implementation.

1 Introduction

There is currently the theory put forward by Nassim Nicholas Taleb, which deals with the nature of the emergence and implementation the information of unexpected events. According to his theory, these phenomena must meet the following criteria [1] :

1. The event is unexpected (for experts);
2. The event has significant consequences;
3. Upon the occurrence, in a retrospective, the event can be given a rational explanation, as if it was expected.

In his theory Nassim Taleb suggests that humanity is not able to successfully predict their future and the confidence in their knowledge advances ourselves and knowledge produces the phenomenon of “over-reliance” [1].

Since in the real world, there are cause-and-effect relationships, then we think we can not say that this problem has no solution, and in our present work, we denote some possible ways to solve it [2]. When creating a news event forecasting model it requires a mathematical tool that would allow formalization the nature of the data and bringing them to a common measurement scale [3]. Obviously, it is impossible to carry out calculation operations in one model, for example, estimates the linguistic and metric scale values without displaying procedures on the formal dimensionless set [4].

The essence of the proposed approach for the prediction of news events is as follows:

1. On the basis of existing methods of mathematical linguistics the description of information space can be formalized in the form of vectors, presenting a set of texts on natural languages. Thus, it is supposed to solve the problem of the heterogeneity of the data and units of various processes settings (measurement scales deliberately roughened and become linguistic, but all the data are formalized in a unified manner).

2. On the natural language it is possible to describe and provided a vector representation of interest (intended) news event for which it will be held forecasting of its implementation.

3. Taking into account that information space is a reflection of the real world in which there is a cause - effect relationships between events, it is possible to assume the implementation of the law of conservation of the information space. And as a hypothesis to test it is

* Corresponding author: zhukovdm@yandex.ru

possible put forward the idea that existing in the information space formalized textual knowledge can shape the image of the projected events of interest.

2 A model the stochastic dynamics of transitions between states in the information space

2.1. Contact meaningful clusters of information space with a predicted event

Take a collection of text documents. Using the methods of mathematical linguistics, they will create a vector representation of the information space. Carry out the clustering of semantic groups at a time t. We define the vectors $(z_1, z_2, ..., z_k, z_j, ...)$, defining the position of the centers of the clusters in a given time. Then spend a textual description of the forecast news events and set its vector X_{bs}.

As a hypothesis, we assume that the information space already has some data on projected event, and we can assume that there should be and the evolution of existing semantic groups of news events in the event that we predict. In the description of evolution in our opinion the most acceptable is the use of the parameters used in informational search whilst determining the relevancy of search requests: finding the distance between the vectors and determination of cosine of the angle between them.
We choose as the evolutionary parameter the finding the projection of x_j vectors defining the position of the centers of the information clusters at any given time $z_1, z_2, ..., z_k, z_j$, in the direction of the vector X_{bs}, which determines the appearance of the predicted events. Each of the projections x_k is defined as the product of the corresponding vector z_k and cosine of the angle between the directions of the vectors z_k and X_{bs} $(x_k = z_k \cdot \cos(\alpha_k))$, that is, in fact, we use the cosine measure, adopted in informational search.

After a certain period of time (called an interval measurement T) values of the vectors defining the position of the information centers of clusters vary to some values Δ_j (j — represents the considered vector). To illustrate, for example, vector z_{01} and z_{02} determine the position of the centers of clusters of news in the information space at a time t, and the vector z_1 and z_2 z_2 – after a time interval T (at time $t + \tau$). In these cases, $\Delta_1 = z_1 - z_{01}$ and $\Delta_2 = z_2 - z_{02}$. Similarly, they determined the changes for positions of the centers of all clusters in the information space at a time τ.

The values x_{01}, x_{02}, x_1 and x_2 are to assign values corresponding to the projections of the vectors defining the position of the news centers of the clusters 1 and 2, in the direction of the vector of predicted events, at times t and $t + \tau$. Note that some values of x_k projections may be larger than the preceding x0k values (for the same group of news data), and some smaller; in the information space of two trends of behavior coexist. One of the projections is to increase the values, the other is to reduce that must be considered in the framework of the developed model. We introduce for any time the concept of the average value $< x_t >$ of all values of the projection vectors that define the position of the centers of clusters of news in the information space on the direction of the axis of the predicted events. For a group of K news cluster at a time t the average value $< x_t >$ is defined as: $< x_t >= (\Sigma x_{t,j}) / K$ (the summation is from j = 1 to K) where $x_{t,j}$ denote the corresponding values of the vectors of the projections, defining the position of the news centers of clusters, on the direction of the axis of the predicted events in the time t. After a period of time τ: $< x_{t+\tau} >= (\Sigma x_{t+\tau,j}) / K$ (the summation is from j = 1 to K). To take account of trends to increase and decrease the values of the projection of the vectors defining the position of the news centers of clusters, on the direction of the axis of the predicted events it can proceed as follows. Based on the analysis x_t the values, j and $x_{t+\tau,j}$ we divide the set of x_t, j into two subgroups, one $(x_{t,j})_{\text{I}}$ will have all the news clusters for which over the time interval τ there was a decrease the quantities of values of the projections $x_{t+\tau,j}$ (denote the number of such clusters as an R), and the second $(x_{t,j})_{\text{II}}$ – increase (denote the number of such clusters as the K–R), and find for each of these the average values $(< (x_{t,j})_{\text{I}} >= (\Sigma (x_{t,j})_{\text{I}} / R$ (the summation is from j =1 to R) and $< (x_{t,j})_{\text{II}} >= (\Sigma (x_{t,j})_{\text{II}} / (K - R))$ projections of vectors defining the position of the centers of these news clusters. Further, we offer the following approach to accounting trends to increase and decrease the values of the projection of the vectors defining the position of the news centers of clusters, on the direction of the axis of the predicted events. Since taking into account the trend it makes sense to talk about the average values, then we will consider the transition to the time interval τ to the point $< x_{t+\tau} >$ of point $< (x_{t,j})_{\text{I}} >$, which is on the axis of events forecasting to the right of $< x_{t+\tau} >$ and the point $< (x_{t,j})_{\text{II}} >$ which is located to the left of $< x_{t+\tau} >$. In themselves the transitions are random events, and their size can be determined as follows: $\xi_t =< (x_{t,j})_{\text{I}} > - < x_{t+\tau} >$ and $\varepsilon_t =< x_{t+\tau} > - < (x_{t,j})_{\text{II}} >$. After the next step, we determine the new values of τ $\xi_{t+\tau}$ and $\varepsilon_{t+\tau}$: $\xi_{t+\tau} =< (x_{t+\tau,j})_I > - < x_{t+2\tau} >$ and $\varepsilon_{t+\tau} =< x_{t+2\tau} > - < (x_{t+\tau,j})_{\text{II}} >$, etc. At any stage n values $\xi_{t+n\tau}$ and $\varepsilon_{t+n\tau}$ may have different random (or almost random) values. Therefore, it is necessary to detect any in their behavior characteristic features (for example, the dependences $\xi_{t+k\tau}$ and $\varepsilon_{t+k\tau}$ from time to time may have self-similarity at not random behavior) or, if they have the characteristics of a uniform distribution, it is possible their averaging over a sufficiently long time period of observation and usage in the model of the average of random values

$\xi = \Sigma\xi_{t+k\tau} / N$ and $\varepsilon = \Sigma\varepsilon_{t+k\tau} / N$ (the summation is from k = 1 to N), where N – number of steps (time intervals τ) monitoring). Note that depending on the values $\xi_{t+k\tau}$ and $\varepsilon_{t+k\tau}$ from time to time can obey a certain distribution law and then its parameters can be set according to obtained data.

2.2. The Construction of difference schemes of transition probabilities between states

In the developed model of processes taking place in the information space, the value of the average value of the current state of vectors projections xi, characterizing the position of the news centers of clusters in the information area on axis of projecting event can randomly increase due to the fact that the value of ε greater than the value ξ at each step (or several successive steps) or randomly decrease if the value is less than the value of ε £. Eventually, state xi would be near the threshold of projected event that equal to vector value X_{bs}.

All set of states will be denoted by X. The state observed at time t can be denoted as $x_i (x_i \in X)$. The observed state is determined by economic and political processes taking place in society. In addition, we introduce a time interval τ, for which is possible to change the state x_i. In this case, any of the current value of time $t = h\,\tau$, where h — the transition step number between states (the transition between states is becoming a quasi-continuous infinitesimal time interval τ), h = 0, 1, …, N. The current state x_i at step h, after the transition to step $h+1$ can occur due to accidental factors increase by a certain amount ε, or decrease by an amount ξ, and respectively be equal to $x_i + \varepsilon$, or $x_i - \xi$. We introduce the concept of probability of the system stated in a particular condition. Suppose that after a certain number of steps h about the described system it can be said that:

- $P(x-\varepsilon,\ h)$ – the probability that it is in a state $(x - \varepsilon)$;
- $P(x,\ h)$ – the probability that it is in a state x;
- $P(x + \xi,\ h)$ – the probability that it is in a state $(x + \xi)$.

After each step, the state x_i (hereinafter index i can be omitted for brevity) may vary by an amount ε or ξ.The probability P $(x, h + 1)$ – that at the next $(h + 1)$ step, the system (or process) in a state x will be equal as follows (see Fig. 1.):

$$P(x,h+1) = P(x-\varepsilon,\ h) + P(x+\xi,\ h) - P(x,\ h). \quad (1)$$

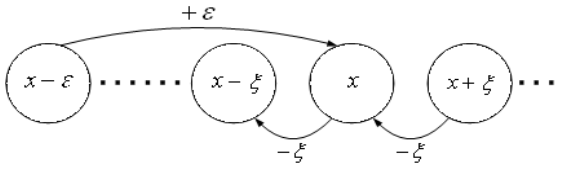

Fig.1. Diagram of the possible transitions between states of the system (or process) to h + 1 step.

Let us explain the expression (1) and shown in Figure 1 the scheme. The probability of a transition in the state x in step h $P(x, h+1)$ is the sum of the probabilities of the transitions in this state of the states $(x - \varepsilon) - P(x - \varepsilon, h)$, and $(x + \xi) - P(x + \xi, h)$ in which the system is located on the step h, net of the transition probability ($P(x, h)$) system from the state x (where she was on the step h) to any other state in the h + 1 step. In this case, we will assume that the transitions themselves are carried out with probability equal to 1.In an actual situation it may remain the memory of the previous state. Therefore, the proposed model must take it into account. We define the probability $P(x–\varepsilon, h)$, $P(x + \xi, h)$ and $P(x, h)$ through the states at the h–1 step. Similarly, the scheme shown in Figure 1is possible to compose schema of the corresponding transitions and write as follows:

$$P(x-\varepsilon,h) = P(x-2\varepsilon,h-1) + P(x-\varepsilon+\xi,h-1) - P(x-\varepsilon,h-1) \quad (2)$$

$$P(x+\xi,h) = P(x+\xi-\varepsilon,h-1) + P(x+2\xi,h-1) - P(x+\xi,h-1) \quad (3)$$

$$P(x,h) = P(x-\varepsilon,h-1) + P(x+\xi,h-1) - P(x,h-1) \quad (4)$$

Having substituted (2), (3) and (4) into the equation (1) and taking into account that $t = h\tau$, where t – time of the process, h – step number, τ – duration of a single step, we move of h to t, and will hold corresponding Taylor series expansion and having limited only by second derivatives, we write:

$$\partial P(x,t)/\partial t = a \cdot \frac{\partial^2 P(x,t)}{\partial x^2} - b \cdot \frac{\partial P(x,t)}{\partial x} - \tau \cdot \frac{\partial^2 P(x,t)}{\partial t^2} \quad (5)$$

where $a = (\varepsilon^2 - \varepsilon\xi + \xi^2)/\tau$; $b = (\varepsilon - \xi)/\tau$.

A member of equations of the type $\partial P(x,t)/\partial x -$ describes an orderly transition to a state when it increases $(\varepsilon > \xi)$, or when it decreases $(\varepsilon < \xi)$; a member of equations of the type $\partial^2 P(x,t)/\partial x^2 -$ describes random change of state (uncertainty). A member of equations of the type $\partial P(x,t)/\partial t -$ can be defined as the rate of overall change in condition of the system over time; a member of equations of the type $\partial^2 P(x,t)/\partial t^2 -$ describes the process by which the states themselves become sources of occurrence of other conditions (self-organization and acceleration as the ordered $(\partial P(x,t)/\partial x)$ and random $(\partial^2 P(x,t)/\partial x^2)$ transitions). From the point of view of the applicability of the model area it should be considered a limitation on the coefficient $a = (\varepsilon^2 - \varepsilon\xi + \xi^2)/\tau$ of the second derivative in x, which takes into account the possibility of accidental changes in the state. It should satisfy the condition $(\varepsilon^2 - \varepsilon\xi + \xi^2) \geq (l - x_0)^2$, the meaning of which is that the transition from the initial state x_0 to achieve the threshold events l may not occur in less than one time step τ. If $(\varepsilon^2 - \varepsilon\xi + \xi^2) < (l - x_0)^2$, the system moves through the door events to achieve in one step.

2.3 The formulation and solution of the problem

Assuming that the function $P(x, t)$ continuous, is possible to go on the probability $P(x, t)$ (equation (5)) to the probability density $\rho(x, t) = dP(x,t)/dx$ and formulate the boundary value problem, the solution of which will describe the process of transition between the states.

$$\frac{\partial\rho(x,t)}{\partial t} = a\cdot\frac{\partial^2\rho(x,t)}{\partial x^2} - b\cdot\frac{\partial\rho(x,t)}{\partial x} - \tau\cdot\frac{\partial^2\rho(x,t)}{\partial t^2} \quad (6)$$

The first boundary condition. The first boundary condition will choose for the following reasons: the state $x = 0$ defines a complete lack of any kind of the processes, which occur in the information space, with their corresponding measured parameters. The actual probability of detection such a condition can be different from zero (though it should be close to zero), but the probability density, which determines the flow in a state $x = 0$, must be set equal to 0 (state of the system cannot go to negative values (it is realized reflection condition), i.e.:

$$\rho(x,t)\big|_{x=0} = 0$$

The second boundary condition. Let us consider the state of the information space with a value of the vector located near the boundary of possible values of its states, denote this limit value as a possible states L. The actual probability of detection such a state will be different from 0.

However, the probability density, which determines the state of flow in the $x = L$, must be set equal to 0 (state of the system cannot go to the range of values greater than the maximum possible value (it is realized the condition of reflection from the boundary)), i.e.:

$$\rho(x,t)\big|_{x=L} = 0$$

Since at time $t = 0$, the state of the system may already be set to a certain value x_0, the initial condition is given in the form:

$$\rho(x,t)\big|_{t=0} = \delta(x - x_0)$$

Since the initial condition contains a delta function, the solution $\rho(x,t)$ is split into two regions at $x > x_0$ and $x \le x_0$. Using the methods of operational calculus for the probability density $\rho_1(x,t)$ and $\rho_2(x, t)$ detection condition in one of the values in the interval from 0 to L, you can get the following system of equations.

When $x > x_0$

$$\rho_1(x,t) = A(x,t)\cdot\Sigma(-1)^n\cdot B(x,t)\cdot\sin\left(\frac{\pi n x_0}{L}\right)\cdot\sin\left(\frac{\pi n(L-x)}{L}\right);$$

When $x \le x_0$

$$\rho_2(x,t) = A(x,t)\cdot\Sigma(-1)^n\cdot B(x,t)\cdot\sin\left(\frac{\pi n(L-x_0)}{L}\right)\cdot\sin\left(\frac{\pi n x}{L}\right),$$

where

$$A(x,t) = -(2\exp(-t\cdot k\cdot(x-x_0)/2\tau))/L,$$

$$k = (\varepsilon - \xi)/(2(\varepsilon^2 - \varepsilon\cdot\xi + \xi^2)),$$

$$B(n,t) = ch(t\cdot(k\varepsilon\xi/(2(\varepsilon-\xi)) - \pi^2 n^2(\varepsilon-\xi)/(2kL^2)^{1/2}/\tau)$$

Amounts are calculated by n = 1 to ∞. If we calculate the integral *P (l, t):*

$$P(l,t) = \int\rho_2(x,t)\,dx + \int\rho_1(x,t)\,dx \quad (7)$$

where the first integral is calculated between 0 and x_0, and the second from x0 to l, then the function $P(l, t)$ will set the probability that the system state at time t time is in the interval from 0 to l ($l = X_{bs}$), i.e., the I threshold event will not be achieved.

Accordingly, the probability $Q(t)$ that the limit event l at time t reached or surpassed can be determined as follows:

$$Q(l,t) = 1 - P(l,t) \quad (8)$$

The analysis shows that $\rho_1(x,t)$ and $p_2(x,t)$ for all values of t and x are non-negative, for the function $Q(l, t)$ as t→∞ following condition is satisfied $Q(l, t)$→1 ($P(l, t)$→0).

2.4 Analysis of the solution of the boundary problem for the dynamics of self-organization of the social conditions in the community

Let us analyze the results. To simulate the process we will assume that the initial (at the start of follow-up) value of condition vector of the system (information space) is $x_0 (x_0 = 0.05$ conditionally accepted the value), the value t_0 assumed to be 1 standard unit of time, $\varepsilon = 0.02$ and $\xi = 0.01$, $l = 2$ – conditionally accepted quantity.

The results of the solution of equation (8) using (7), function $\rho_1(x,t)$, $\rho_2(x,t)$ and a predetermined above mentioned set of parameters and different event thresholds (note that in this case the predicted event is observed during the growth of the value of the vector of system state), selected in the simulation are presented in graphical form in Fig. 2.

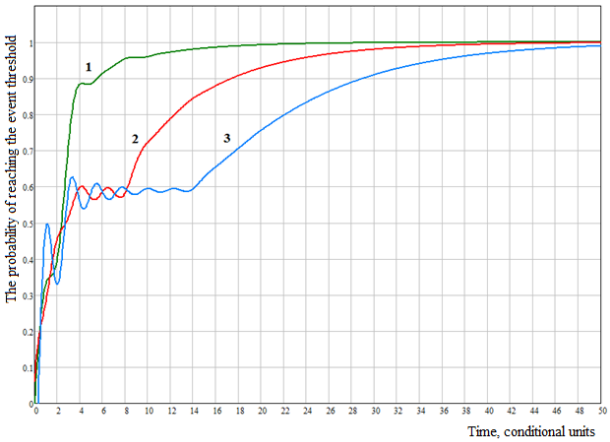

Fig. 2. Graphical representation of the results of modeling to overcome the percolation threshold of negative attitudes in society.

Curve 1 in Figure 1 is constructed for the event threshold equal to 0.1; curve 2 for the magnitude of event threshold equal to 0.2, and curve 3 for the event threshold value of 0.3.

The course of the curves in Figure 1 shows the possibility of the growth of the transition probability threshold of event almost immediately after the start of the process, which is associated with the presence of the

memory about the previous state of the system in a model developed by us, and the possibility of significant self-description of the system due to differential equation of the model member which has the type – $\partial^2\rho(x,t)/\partial t^2$.

Curves 2 and 3 in Fig. 2 show that the nearer value of x_0 of the system state at the start of observation to the threshold event, the faster the transition probability increases (curve 1 is built for event threshold of 0.1, and curve 2 for 0.2, with the same initial value of the system state vector – 0.05). The course of the curves in Fig. 2, shows the possibility of the growth of the transition probability of percolation threshold, almost immediately after the start of the process, this is due to the presence of the memory of the previous state of the system in our model developed, and the possibility of self-description of the system as a result of the account in the differential equation model of the member which has the form $\partial^2\rho(x,t)/\partial t^2$.

The second feature of our proposed model is the possibility of several abrupt changes across the threshold the transition probability event.

The third feature of our proposed model is the existence of phenomena in the wave-like behavior for the value of achieving the threshold event.

The developed model takes into account many of the basic properties of rare events: a time uncertainty of their existence, stochastics with unknown characteristics, and the presence of memory in a system the event occurs, the self-organization of information.

Our model allows us to analyze the possibility of predicting the rare news event in conjunction with the various clusters in the information space. An analysis of the values of times achieving the rare events may allow them to make a fairly accurate prediction for a given level of probability to realize (for example 0.90 or 0.95).

It should be noted that every rare news event has the poorly structured and poorly defined precursor, the appearance of which was not revealed, but they are meaningful indicators in a retrospective to give the usual rationalized definition the phenomenon occurring. The study of self-similarity of information processes may allow us to determine the period or interval of self-similarity, which is important from the prognostic point of view.

3 Conclusions

The proposed model considers the uncertainties in the occurrence of events in the information space, and is not based on the statistical characteristics of a pre-supposed law distribution.

The proposed model shows the possibility of the growth of the transition probability threshold to achieve news event in the information space almost immediately after the start of its process of developing, due to the taking into account the memory of previous states of the system and the possibility of significant self-description due to differential model of the second derivative by time.

The proposed model shows the possibility of abrupt changes for the transition probability through the event threshold and takes into account the presence in her behavior wavelike phenomena.

The developed model allows creating algorithm of analyzing the relationship of information clusters in information space with the possibility of realization of any event, as well as to determine the predicted time of its occurrence.

The work is executed due to financing by the Ministry of Education and Science of the Russian Federation as the competitive part of government tasks of higher education and scientific organizations for the implementation the initiative scientific projects; the number of the project 28.2635.2017 / IF named "Development of models of stochastic self-semistructured information and implementation of memory in predicting the news events array-based natural language texts.

References

1. N.T. Nassim, *The Black Swan: The Impact of the Highly Improbable* (Random House, 2007).
2. J. M. Hofman, A. Sharma, D. J. Watts, Science, **355**. 486 (2017)
3. D. Roy, D. Ganguly, M. Mitra, G. J. Jones, Neu-IR'16 SIGIR Workshop on Neural Information Retrieval, Pisa, Italy (2016)
4. J.R. Bellegarda, C. Monz, Computer Speech & Language, **35**, 163 (2016)