

# *Scientific Bases for Stock Market Fiasco Forecasting Technology with Use of Information Space Entropy*

**Vasily Karasev**

IPME RAS, Intelligent Integrated Systems  
of Automated Design Laboratory  
Saint-Petersburg, Russia

**Ekaterina Karaseva**

State University of Aerospace Instrumentation, Institute of  
Entrepreneurship Technologies, Department of  
Entrepreneurial Information Technologies,  
Saint-Petersburg, Russia

**Abstract** - The article contains a theoretical study and description of general algorithm for predicting a stock market fiasco caused by non-financial and other factors. Market fiasco is considered as non-periodical, sudden and random event which can arise due to the many latent reasons. Methods of technical and fundamental analysis are useless to solve this problem, therefore, the use of systems analysis methods is proposed. The author's idea is the numerical calculation of search queries entropy as a part of global information space. Decrease in the Renyi's entropy, associated with rapid grow search queries, containing key terms from the subject area, indicates the possible stock market fiasco in the near future. This article presents an algorithm for the dynamic calculation of Renyi's entropy, allowing predict rare events which are not reflected in statistical data (or frequency of their realizations is too small). The method and algorithm can be realized in trade information systems and decision-making systems in economic sphere.

**Keywords** - Stock market, uncertainty, risk, forecasting, fiasco, collapse, Renyi entropy, information space, frequency, algorithm, technology.

## 1. INTRODUCTION

The recent events, dramatic for the global economy, highlight a number of systemic problems in the area of management, forecasting, data analysis and decision making. There is an necessity to develop new effective technologies for forecasting and risk management in economic systems that can predict the emergence of "black swans" [1], i.e. rare and unpredictable events which were not reflected in historical data for past years within the foreseeable time horizon.

To mitigate the consequences of the stock market fiasco (to reduce the loss of resources) or try to avoid them we need to predict this undesirable event in advance. Development of information technology for forecasting the market fiasco (market failure, market shock) is very promising project and this allows us to reduce uncertainty [2] and be ready in the future for a sudden market collapse with change in the structure of economic relations.

## II. THE PROBLEM OF MARKET FIASCO FORECASTING BY TRADITIONAL METHODS

There are a large number of tools [3] which allow predict economic parameters, many of them are based on time series analysis [4] and the hypothesis of the stationary time series.

On the charts of prices and returns, we can sometimes observe bursts, price jumps, which are described as random outliers (or market shocks). Stabilized period follows after such extreme change of values, then, the process of changing the values becomes stationary again.

During the market fiasco (collapse, market shock), a surge is observed on the price and yield charts - a rapid increase or decrease of the parameter (price jump), then long transition process is, and then either a process is slowly diminish, or, in most cases, the market comes back to a conditionally stable state and the time series become periodical and stationary again.

Such surge and the following stabilization period are modeled by the logistic curve (sigmoid) [5], which is given by the function:

Online ISSN 2256-070X

<https://doi.org/10.17770/etr2021vol2.6508>

© 2021 Vasily Karasev, Ekaterina Karaseva. Published by Rezekne Academy of Technologies.  
This is an open access article under the [Creative Commons Attribution 4.0 International License](#).

$$f(x) = \frac{Mf(x_0)}{(M-f(x_0))e^{-rx}+f(x_0)}, \quad (1)$$

where:  $e$  is natural logarithm (Euler's number),  $f(x_0)$  is the function value at initial time ( $f(x_0) \neq 0$ ),  $r$  is the logistic growth rate (steepness of the curve),  $M$  is maximal growth value.

We can predict changes in economic parameters during the market fiasco if parameters  $f(x_0)$ ,  $M$ ,  $r$  are correctly determined before. However, we can determine parameters of logistic curve by a posteriori analysis when the market has already collapsed and we know the peak values, amplitude and duration of the surge. For the market fiasco forecasting, such analysis is useless, because each stock market fiasco is caused by particular reasons and has individual evolution scenario. Accordingly, the values of  $r$  and  $M$  are unknown before the collapse.

For predicting of random events, such as a stock market fiasco, which realize due to the influence of unidentified latent factors in time, traditional methods, based on time series analysis, are not suitable.

Fundamental analysis [6] is also useless here, since financial and production indicators are used only while a market fiasco can realize due to many random and unpredictable events, including events of non-financial nature.

### III. ENTROPY AND ANALYSIS OF INFORMATION SEARCH QUERIES SPACE

A complex non-trivial problem can be solved by interdisciplinary approach [7], based on the methods of system analysis, applied to large amount of initial data (BigData).

The authors' idea is to develop a technology for predicting the stock market fiasco (collapse) based on the numerical calculation of the information space entropy, and, more specifically, search queries entropy.

Information space is a set of databases, which includes publications from accredited mass media, the content of information sites, blogosphere and social networks.

Entropy is a measure of the uncertainty of a system [8].

In our interpretation, the information space entropy describes the uncertainty of the content of publications in the media and the blogosphere, and the variability of their themes.

In a stable period with low volatility, the media inform us about a variety of topics, variability and entropy are high. If large problems appear and the volatility stock market increases, media is paid more attention to these problems, therefore, the media content has similar topics, variability decreases, entropy also decreases. The increased frequency of same terms in different media sources makes the content more predictable.

The methodology for calculation the index of economic policy uncertainty EPU [9, 10, 11] is based on the frequency of certain key terms. The EPU is a set of indices that are important factor influencing on choice an investment strategy [11]. The set includes the world index, countries and regional indexes. Also, local indexes can be defined, for example, for a specific region. EPU values are also lagging in time because they are determined by a posteriori analysis of press publications, when undesirable event have realized. Since the world EPU index are calculated monthly, the delay in the value is about a month.

The information space structure can be presented, generally, as two fundamental units: information field and information flows. Information field is a set of databases (data storages) but information flows are set of inquiries and set of corresponding responses.

We can estimate the probability of a stock market fiasco a priori by analysis of search queries in the information space. In [12] authors studied the relationship between the sentiment of messages in the Internet (publications in the media and social networks) and changes in the yields of specific stocks. Data about 3238 stocks was used from 2005 to 2009. The authors observed the relationship between the number of publications (and the number of discussions) and the trading volume but stocks of different companies demonstrate different sensitivity to the tone of publications (airlines are more sensitive to the tone of the media, IT companies are less sensitive). The authors emphasize, if we apply strategies based on the analysis of publications in the media, blogs or social networks, these strategies provide the best results on short forecast horizons because the market quickly responds to news [13].

The Google Domestic Trends tool allows predict market fiasco using analysis of popular search queries [14].

Forecasting other events, leading to the market fiasco and caused by non-financial reasons (for example, the spread of a viral infection) is also possible for short horizons by analysis of popular queries using contextual data analysis [15]. For example, the Google Flu Trends service allows predict influenza epidemic in short horizons (about two weeks before the epidemic) by analyzing search queries.

These tools use a single principle for analysis the state of the subject area. For each area, a set of appropriate keywords is identified, then, a chart is plotted by the number of search queries. The chart is used for analysis of the public interest in a particular area and we can assume how the market state will change corresponding to public interest.

Nevertheless, there is some skepticism among specialists about the practical application of these tools [16].

Skepticism is caused by inaccurate predictions, which, in our opinion, are explained by the fact that the above

tools analyze frequencies of separate requests, while a rapid increase in the frequency of requests, containing keywords only, cannot indicate an impending event without a doubt.

To overcome this problem, we propose to take into account not only increasing frequency of queries, containing key terms, but also changing entropy in search queries space. There is always some nonempty set of terms, which may indicate future fiasco and have implicit context sense, but not included in set of key terms. Frequencies of such terms can also correlate with frequencies of key terms. Calculation of entropy allow us to take into account the influence of these terms and significance of secondary factors [17] that lead to the predicted event. This is important part of the information space analysis.

Search queries are a part of the information space and can be used to estimate the likely future changes in the entire space. In other words, in practice, search queries are a good proactive indicator of the system's state (both the stock market and society before a virus pandemic). The search query space also has entropy, a change in entropy can inform us about future change in the information space entropy.

We propose to predict the stock market fiasco based on the dynamic calculation of the entropy of search queries: a decrease in entropy indicates an imminent collapse of the market, an increase inform about future stable state. Note, in thermodynamics we have a similar situation: in a nonequilibrium (irreversible) process (all real physical processes are irreversible) - the closer the system's state to equilibrium (stability), the greater the entropy [18]. A stable state corresponds to maximum entropy.

Let's denote a set of search terms as  $\Omega$ . Each search query is described by the presence (or absence) key terms and a region parameter.

To calculate the entropy, we introduce a finite set of key terms  $ST = \{s_1, s_2, \dots, s_k\}$ , associated with a finite set  $\Psi$  containing negative factors with various nature. Let us denote by  $Z = \{z_1, z_2, \dots, z_n\}$  the set of all terms, which frequency in search queries exceeds a certain set frequency threshold  $w_{ad}$ , and  $ST \subseteq Z$ .

Let suppose, at time  $t$  each of the terms  $s_i, i = 1, \dots, k$  has a frequency of occurrence in messages  $\omega_i, i = 1, \dots, k$ . If we consider the search queries space at sequential times  $t_1, t_2, \dots, t_m$ , then at each time  $t_i$  we will get frequency vector  $\{\omega_{1t_1}, \omega_{2t_1}, \dots, \omega_{kt_1}\}, \{\omega_{1t_2}, \omega_{2t_2}, \dots, \omega_{kt_2}\}, \dots, \{\omega_{1t_m}, \omega_{2t_m}, \dots, \omega_{kt_m}\}$ .

The volume of queries in time  $t$  around the world is huge, so, we can assume, frequencies  $\omega_i, i = 1, \dots, k$ , exceeding frequency threshold  $w_{ad}$ , are close to the objective values of probabilities  $p_i$  of realizations the terms  $s_i$  at time  $t$  in a set of messages. The vector of probabilities  $\{p_1, p_2, \dots, p_k\}$  corresponds to the vector of terms  $\{s_1, s_2, \dots, s_k\}$ .

The random variable  $S$  takes the value  $s_i, i = 1, \dots, k$  from the set  $\{s_1, s_2, \dots, s_k\}$  with probability  $p_i, i = 1, \dots, k$ .

To analyze the query space, we will use the generalized Renyi entropy [19, 20], which describes the quantitative diversity of the randomness of queries. The Renyi entropy is calculated by the formula:

$$H_a(S) = \frac{1}{1-a} \cdot \log(\sum_{i=1}^n p_i^a), \quad (2)$$

where  $p_i = P(S = s_i)$  is the probability the discrete random variable  $S$  will be equal to the corresponding possible value of  $s_i$ ;  $n$  is the total number of different possible values of the random variable  $S$ , that is, the number of given terms  $z_i$  for which  $\omega_i > w_{ad}, i = 1, \dots, n$ ;  $a$  - a given real number that meets the requirements  $a \geq 0, a \neq 1$ .

Theoretically, when the situation is stable and the market fiasco is not expected in the near future, we have a uniform distribution and the probabilities are so close to each other that they can be considered as equal, i.e.  $p_1 = p_2 = \dots = p_n = 1/n$ , the Renyi entropy is  $H_a(S) = \log n$ . Otherwise, the entropy starts to decrease with increasing  $a$ . Renyi entropy is used in ecology and statistics as indices of diversity, as well as in quantum information theory, where Renyi entropy is a measure of complexity.

In our problem, the Renyi entropy formula is needed for special cases:

1.  $a = 0$ , we obtain  $H_a(S) = \log n$ , where  $n$  is the cardinality of the area of possible values of a finite random variable  $S$ , i.e.  $n$  is the number of different terms belonging to the set of possible values  $Z$ .
2.  $a = 1$ , in this case  $H_a(S)$ , according to L'Hôpital's rule, reduces to the Shannon's information entropy:

$$H_1(S) = - \sum_{i=1}^n p_i \cdot \ln(p_i) \quad (3)$$

The algorithm for dynamic analysis of the information space of queries looks like a regular sequence of following operations:

1. Define the set  $\Psi$  with all possible factors that can cause a sudden fiasco (collapse) of the market;
2. Define the set  $ST = \{s_1, s_2, \dots, s_k\}$ , containing key terms related to the set  $\Psi$ ;
3. Define the set  $Z = \{z_1, z_2, \dots, z_n\}$ , containing a set of terms, having a high correlation with key terms, the frequency of these terms in search queries exceeds the specified threshold  $w_{ad}$  in current time  $t_m$ ;

4. Determine frequencies  $\omega_i$ , and probabilities  $p_i$  for the terms  $s_i$ ,  $i = 1, \dots, k$ , using the tools of semantic analysis of queries;

5. Calculate the numerical value of the entropy  $H_a(S)$  by the formula (2) with the value  $a = 0$ , or  $a = 1$ , depending on the values of  $p_i$ ,  $i = 1, \dots, k$ .

If we observe a decrease in entropy in comparison with value in the previous time  $t_{m-1}$ , then this is a signal about imminent market fiasco.

#### IV. CONCLUSION

The proposed method has good perspectives for practical application but there are a number of problems that still need to be researched:

1. Time horizon of forecasting. There are no exact numerical estimations, only empirical data obtained from the results of using GoogleTrends tool;

2. Recommendations for the choice  $w_{ad}$  value can be formulated only after testing of the algorithm. To this moment, authors can only recommend to choose this value by the method of iterative approximation, based on the results of the preliminary semantic analysis of queries. The correct choice of the value will reduce the dimension of the task without losing the information and completeness of the set. We need more research to provide accurate recommendations.

3. We don't know whether the value of the change in entropy indicates the scale of the coming crisis. If we take into account the region parameter, then a decrease in the the search query space entropy in a particular region can predict financial problems only in this region.

4. Determination a set of key terms  $ST$  is a non-trivial problem. The solution will require data processing and frequency analysis for large number of queries.

Nevertheless, due to large cardinality of the set  $Z$ , the stated theoretical developments and the algorithm can be used in artificial intelligence systems for risk management that support the technology of processing and analyzing big data.

The proposed algorithm can be applied in complex technical systems also, if functioning processes in large industrial corporations are united into the integrated information space within the framework, offered by "Industry 4.0" concept [21]. In this case, a change in entropy of large data arrays may indicate about accumulating problems (the "snowball" effect), which will eventually find a way to be realized in the production and management system of a corporation through uncontrolled functionally weak links (the so-called "Swiss cheese holes" effect [22]) and will lead to the collapse or systemic failure in the corporation's functioning processes.

#### REFERENCES

- [1] N.N. Taleb, *The Black Swan: The Impact of the Highly Improbable*. Random House, New York, 2007, 400 p.
- [2] S.R. Baker, N. Bloom, S.J. Davis, S.J. Terry, "COVID-Induced Economic Uncertainty". *NBER Working Paper Series* // National Bureau Of Economic Research, Cambridge, 2020, 16 p.
- [3] *Mezhdunarodnaya praktika prognozirovaniya mirovyyh cen na finansovyh rynkah (syr'e, akcii, kursyvalyut)* / pod red. Ya.M. Mirkina. — M.: Magistr, 2014, 456 p. (In Russian)
- [4] R.H. Shumway, D.S. Stoffer, *Time Series Analysis and Its Applications With R Examples*. Fourth Edition, Springer, 2016, 558 p.
- [5] R.A. Ramos "Logistic Function As a Forecasting Model: It's Application To Business And Economics". *International Journal of Engineering and Applied Science*, March 2013, Vol. 2, N. 3, pp. 29 – 36.
- [6] B. Graham., D. Dodd, *Security Analysis: Sixth Edition*, Foreword by Warren Buffett, McGraw-Hill Irwin, 2008, 766 p.
- [7] C.S. Wasson, *System Engineering Analysis, Design, and Development. Concepts, Principles, and Practices*. John Wiley & Sons Inc., Hoboken, New Jersey, Canada, 2016, 846 p.
- [8] D.N. Zubarev, V.G. Morozov, *Entropiya // Fizicheskaya enciklopediya*: [v 5 t.] / Gl. red. A. M. Prohorov. — M.: Sovetskaya enciklopediya (t. 1-2); Bol'shaya Rossijskaya enciklopediya (t. 3-5), 1988-1999. (In Russian)
- [9] "Economic policy uncertainty". [Online]. Available: <http://www.policyuncertainty.com/> [Accessed: April 05, 2020];
- [10] "Global EPU Index through May 2017". [Online]. Available: [https://policyuncertainty.com/media/Global\\_Annotated\\_Series.pdf](https://policyuncertainty.com/media/Global_Annotated_Series.pdf) f – [Accessed: April 05, 2020];
- [11] Ch. Syueczyun', *Indeks neopredelenosti ekonomicheskoy politiki I volatil'nost' fondovogo rynka Kitaya primentel'no k Rossii* / CH. Syueczyun', M. Tyan', SH. YAn' // Innovacii i investicii. – 2019. – №9. – pp. 99-104; (In Russian)
- [12] W. Zhang, S. Skiena, *Financial Analysis Using News Data*. Stony Brook University, Stony Brook, NY, USA, 2008, 71 p.
- [13] S.J. Grisafi, *Market Dynamics: The Mechanics of Financial Engineering*. CreateSpace Independent Publishing Platform, 2013, 218 p.
- [14] C. Curmea, T. Preisb, H.E. Stanleya, H.S. Moat, "Quantifying the semantics of search behavior before stock market moves", *PNAS*, August 12, 2014, vol. 111, No. 32, pp. 11600–11605.
- [15] A.V. Semyonova, M.V. Korsunskaya, *Kontent-analiz SMI: problem I opyt primeneniya* / Pod red. V. A. Mansurova. — M.: Institut sociologii RAN, 2010, 324 p. (In Russian)
- [16] "Eksperiment: Ispol'zovanie Google Trends dlya prognozirovaniya obvalov fondovogo rynka" in *Blog kompanii ITI Capital*. [Online]. Available: <https://habr.com/ru/company/iticapital/blog/279021/> [Accessed February 07, 2020, 18.39 p.m.] (In Russian)
- [17] M. Gladuell, *Perelomnyj moment. Kak neznachitel'nye izmeneniya privodyat k global'nyim peremenam*. Izdatel'stvo: Al'pinaPabliher, 2015, 374 p. (In Russian)
- [18] P. Glensdorf, I. Prigozhin, *Termodinamicheskaya teoriya struktury, ustojchivosti i fluktuacij*. M., 1973, 280 p. (In Russian)
- [19] A. Renyi, *On Measures of Entropy and Information* / A. Renyi // Proc. Fourth Berkeley Symposium - V.1 - Berkeley. Calif.: University of California Press, 1961, pp. 547-561.
- [20] O.L. Korolev, M.Yu. Kussy, A.V. Sigal, *Primenenie entropii pri modelirovanii processov prinyatiya reshenij v ekonomike*. Pod red. A.V. Sigala. Simferopol': Izdatel'stvo «ODZHAK», 2013, 148 p. (In Russian)
- [21] "Chetvertaya promyshlennaya revolyuciya: Celevye orientiry razvitiya promyshlennyh tekhnologij i innovacij" in *Informacionnyj dokument, Vsemirnyj ekonomicheskij forum*, 2019, 49 p. (In Russian)
- [22] E. Foer, "Metafora katastrofy: kak upravlyat' riskami pri pomoshchi shvejcarskogo syra". *Theory & Practice*, 2013. [Online]. Available: <https://theoryandpractice.ru/posts/7427-metafora-katastrofy-kak-upravlyat-riskami-pri-pomoshchi-shveysarskogo-syra> [Accessed: May 05, 2020, 23.37 p.m.] (In Russian)