UDC 681.513.7; 519.688

Y. Protsiuk, O. Gavrylenko

APPROACHES TO THE SOLUTION TO THE PROBLEM OF NEWS-BASED EVENTS FORECASTING

Abstract: An overview of the areas of application of approaches and methods of forecasting events based on past events. The substantiation of urgency of a theme is given and possibilities concerning application of results of work are resulted. Requirements for incoming news regarding their quality are defined. It is noted that there are four key criteria for the quality of the media, which are often two-component, namely: the relevance of news, providing the context in which the event, compliance with professional standards and a variety of materials. The key stages of working with data in order to obtain knowledge from them for forecasting events are identified. These include pre-processing of data (reduction to a standardized view that will understand and be able to process the algorithm), their analysis and the forecasting process itself. The spheres of application of associative series and Markov processes for search of causal relations, and time series for definition of the period of occurrence of an event with the set probability are specified.

Keywords: event forecasting, data mining, news quality, computer linguistics, time series, associative rules.

Introduction

The media have long ceased to be just a source of news about world events. It is impossible to deny how much the media influences the human mind and the course of events in general. Increasingly, you can see how the news becomes a harbinger of certain events, as if calling them. Therefore, there is a question of building a model that will help predict future events based on information about the past. Gaining knowledge of natural languages is an important and little-studied issue of data mining. Creating quality models can solve the global problem of finding hidden patterns that can predict possible future news events and create models of influence on various economic and social processes. An example of predicted phenomena may be the emergence of economic crisis or social processes.

Some knowledge of the future will allow people or especially the leadership to take measures to mitigate or avoid adverse events or, conversely, to create certain trends in the economy. All this can potentially affect the fate of mankind. In any case, the identified patterns are a field for scientists from different fields, not just computer scientists or sociologists. The analysis of texts and the extraction of data from them for use in forecasting models can be based on known methods of computational linguistics and text mining.

Analysis of recent publications

In general, the analysis of news to predict future events is a topic that has been little studied due to its complexity. In the study [1] with the help of natural language processing methods studied the peculiarities of the emergence of interconnected sensational news based on text analysis. The article examines the patterns of occurrence of pairs of events that occurred in the news space and how to predict that the second event will occur after the first. Computational linguistics methods have been used to find cause-and-effect relationships between events from their test descriptions.

Research [2] is devoted to the identification of causal links between events on social networks to predict the tone of the event and the time between the occurrence of different events. First, messages are selected over a period of time, from which the keywords used to determine the tone of the message - positive, negative or neutral - are selected. To determine the tone of words, a classifier is used, which is studied on the basis of the method of reference vectors. In the future, causal relationships are built between keywords, using the method of associative rules, which creates rules of the form "if" from the data. The final step is to predict events using temporal analysis of messages and the calculation of causation.

Research [3] is based on linguistic analysis and statistical modeling of tweets to automatically identify topics discussed in large cities. To single out topics, it is recommended to use thematic modeling, for which the text of the tweets was divided into special tokens using a language tokenizer and a partial tag. In this case, emoticons were considered as separate tokens that carry a certain content load. To model the topic, semantic content is also analyzed, which describes the emotional state of the author of the tweet.

The problem of the influence of news headlines on the behavior of investors and changes in the financial market is covered in [4]. A model based on prudent associative rules determines whether news is important enough for investors. When learning from real-world data, the weighted associative rules algorithm finds terms that appear frequently and simultaneously in news headlines. The term appears in the headlines several times a day for a certain number of days. And the severity of the impact of the term is determined by how much the share price has changed over the period, taking into account the frequency of use of this term. These scales allow you to determine whether certain terms affect the results of trade.

Some researches propose methods to solve the problem of identifying events that are a harbinger of future events and identifies future events. According to the collection of streaming news from open sources, an embedded approach has been developed to predict significant public events and protests. The strengths of this approach are proved by empirical assessment, which consists in filtering potential precursors in order to qualitatively predict the signs of events of joint riots and in predicting the instance of an event ahead of time [5].

 ISSN 1560-8956

The authors of the study [6] present a model for predicting fatal accidents and natural disasters. The authors have collected text messages from Google about disasters. The resulting text documents were processed using computational linguistics methods and erroneous results were eliminated using a trained naive Bayesian classifier. After data collection, semantic clustering of this data was performed. The transition matrix was built from the keywords used in data collection. The observation matrix was built from grouped events. Both matrices were fed to the input of a hidden Markov model for prediction. To predict a new event with the specified topic, it is necessary to develop a model of origin based on its time series, and then find the density function of the distribution of its parameters. When forecasting, the main problem with time series analysis and modeling is that there is only one process implementation at a time (one statistical sample, one time series sample already implemented) that needs to be used to make a forecast for the future. Regardless of the tools used in the analysis method: statistical models, neural networks or fuzzy logic models, the nonstationary time series is divided into certain sections, where it is quasi-stationary with its selective distribution function, and there are parts of the series in which transients occur. The duration of the transition process is determined by both the physical changes and the sample size used for statistical analysis. The parameters of the distribution function are determined based on the analysis of data on the time interval of quasi-stationarity. In particular, nonparametric methods help to restore the probability density based on the observed values. In practice, there are two problems: to determine the time interval of quasi-stationarity and to determine the beginning of the transition period with minimal delay.

To predict the upcoming news, it is advisable to consider time dependences in the flow of events and introduce a piecewise constant approximation of their intensity, using the Bayesian approach and Poisson's distribution to describe future events.

General principles of forecasting news events

As a generalization, we can offer such an approach for forecasting news events based on the analysis of the dynamics of events that have already taken place. The first step is to collect a set of text data from news sites. It is recommended to use the news in English first, because there are more convenient models and tools for further processing. Today, in connection with the reduction of expenditures on information publications, the quality of the media is a particularly important topic that directly and indirectly affects politics, economics and culture. Therefore, proven quality measurements are important for assessing the state of media systems. Unfortunately, this process is not easy due to the double hermeneutics of the social sciences. The ways in which sociologists assess the quality of news media constantly interact with the values of the structures that divide society. This makes the quality of news media a dynamic, conditional and contradictory construction. Nevertheless, by identifying the four main components of media

quality (object, ideal, class, and criteria), a fairly clear and adequate media evaluation system can be obtained. The main criteria for the quality of the media are the relevance of news, providing the context in which an event is, compliance with professional standards and a variety of materials (figure 1).

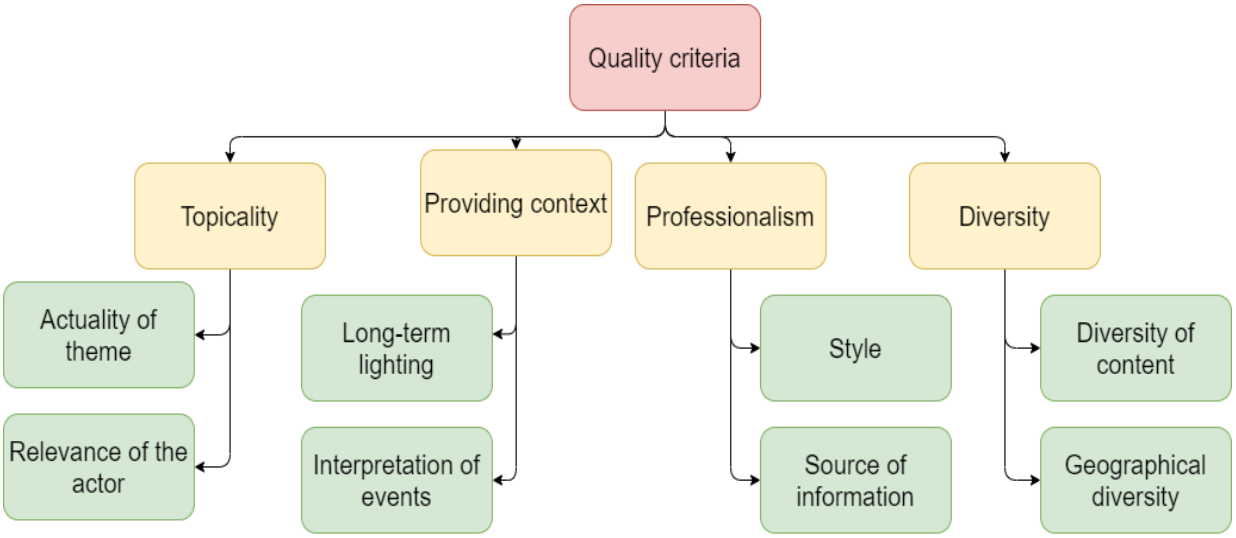

Figure 2. Media quality criteria

With the help of computational linguistics methods, it is necessary to carry out the following preliminary data processing: delete stop words, perform stemming or lematization and vectorization based on existing dictionaries, create a TF-IDF matrix. Then it is necessary to cluster by thematic groups with the date and time of the news [7]. The input data will be considered a vector representation of the textual description of predicted events, which allows you to find the cosine of the angle between the text and the centroids of thematic clusters derived from the news collection. The change in the value of this cosine in time is considered as a wandering point on the segment [0,1], which contains a trap in the threshold point of the event, where the wandering point can get over time. The minimum value of the allowed cosine metric similarity metric should be considered as a trap. It is necessary to pay attention to the probabilistic schemes of transitions between states in the information space. The parameters of the model can be determined on the basis of the analysis of changes in the structure of the clusters of news over time. The location of the cluster centroid vector and the number of thematic messages per day can be considered as a non-stationary time series. The appearance in the news feed of descriptions of events related to a particular topic, over time, can be considered as the formation of a discrete time series (parameter - the frequency of mentions of the event during the day).

Analysis of the dynamic characteristics of a series can be used to predict its change, as well as to calculate the probability of an event occurring over a period of time. Keep in mind that in order to form a time series, you need to solve the problem of selecting from the news feed text

 ISSN 1560-8956

messages related to this topic with the highest accuracy. This will ensure that much of the information is not lost in the formation of the time series, for example, the frequency of the event, which will achieve a more accurate definition of the parameters of the time series and will have no effect on its development. As a result, the predicted event can be formed from a set of events from clusters.

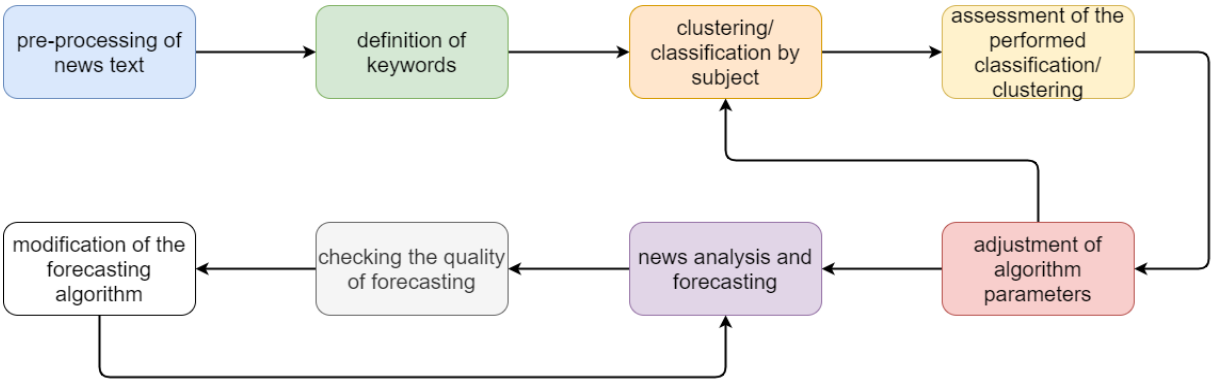

Figure 2. Approximate scheme of work on forecasting events based on news

Conclusions

The problem of event forecasting based on news processing for the previous period is formulated and defined. It is noted that the problem is complex and may involve the use of such mathematical apparatus as statistical analysis, time series, Markov models and processes, associative rules. An overview of publications on the topic, which methods of solution are used more often than others. The disadvantages of these methods and ways to avoid them are identified. A generalized scheme of problem solving is proposed, which provides for pre-processing of text data, classification or clustering of data for their analysis, application of the prediction method (based on Markov processes, time series, associative rules or neural networks).

REFERENCES

1. Discovering and learning sensational episodes of news events / X. Ao et al. Information Systems. 2018. Vol. 78. P. 68-80. DOI: 10.1016/j.is.2018.05.003.

2. Preethi P. G., Uma V., Kumar A. Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction. Procedia Computer Science. 2015. Vol. 48. P. 84-89. DOI: 10.1016/j.procs.2015.04.154.

3. Anastasiu D. C., Tagarelli A., Karypis G. Document Clustering: The Next Frontier. Data Clustering. 2018. P. 305-338. DOI: 10.1201/9781315373515-13.

4. Realization of a news dissemination agent based on weighted association rules and text mining techniques / C. Huang et al. Expert Systems with Applications. 2010. Vol. 37, № 9. P. 6409-6413. DOI: 10.1016/j.eswa.2010.02.078.

5. Modeling Precursors for Event Forecasting via Nested Multi-Instance Learning / Y. Ning et al. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. DOI: 10.1145/2939672.2939802.

6. Singh S., Khatri R. Data Mining based Technique for Natural Event Prediction and Disaster Management. International Journal of Computer Applications. 2016. Vol. 139, № 14. P. 34-39. DOI: 10.5120/ijca2016909102.

7. Zhukov D., Andrianova E., Trifonova O. Stochastic Diffusion Model for Analysis of Dynamics and Forecasting Events in News Feeds. Symmetry. 2021. Vol. 13, № 2. P. 257. DOI: 10.3390/sym13020257.

 ISSN 1560-8956