Peer-to-Peer Networking and Applications (2021) 14:2626–2634
https://doi.org/10.1007/s12083-020-00978-3

Analysis and mathematical modeling of big data processing

Kairat Imanbayev[1] · Bakhtgerey Sinchev[2] · Saulet Sibanbayeva[3] · Axulu Mukhanova[1] · Assel Nurgulzhanova[4] · Nurgali Zaurbekov[5] · Nurbike Zaurbekova[6] · Natalya V. Korolyova[3] · Lyazzat Baibolova[1]

Received: 2 March 2020 / Accepted: 4 August 2020 / Published online: 12 August 2020

Abstract

Big data processing is an urgent and unresolved challenge that originates from the intensive development of information technology. The recent techniques lose their effectiveness rapidly as the volumes of data increase. In this article, we will put down our vision of the basic approaches and models related to problem solving, based on processing large data volumes. This article introduces a two-stage decomposition of a problem, related to assessing management options. The first stage of our original approach implies a semantic analysis of textual information; the second stage is built around finding association rules in a database, processing them via mathematical statistics methods, and converting data and objectives to a vector. We suggest processing the collected news events by a semantic model, which describes their key features and interconnections between them in a specified subject area. The classification-based association rules allow assessing the likelihood of a particular event using a set chain of events. This approach can be applied through the analysis of online news in a specified market segment.

Keywords Big data · Modeling · Identification · Classification · Association rules · Data mining

1 Introduction

Big data and big data processing are currently concerning to both science and business. Big data processing is focused on transaction-oriented, multimedia-intensive, and other tasks [1]. Data volumes are continuously increasing, which requires new or improved software applications to handle them. Therefore, electronic information about objects and related processes represents data gathered by management information systems [2, 3]. To achieve sustainable development of information technologies, a thorough study of innovations and their rational and phased application in various fields including big data management is needed [34]. The challenge of big data management originates from the increase of information and the subsequent rising of data processing requirements [4]. For example, many enterprise resource planning (ERP) systems are focused on providing information about all company's business processes to improve its overall efficiency at a lower cost but these systems do not take technical specifications of the product into account [5]. Normally, this type of data is managed by special management systems designed for specific product criteria and regulated by a general management system [6]. Besides, these systems carry a number of other functions that allow companies to properly manage and make use of product data. For instance, a web application for storing and managing product data in a decentralized repository allows easily distributing and updating product information that is reflected in a web catalog [7]. Apart from the increasing data volume, ERP system faces a challenge of complex information systems, the multiple components of which result in cost-intensive data analysis. The essential function of organizational information system is to create and visualize reports, but converting data from several sources and different

This article is part of the Topical Collection: *Special Issue on Security of Mobile, Peer-to-peer and Pervasive Services in the Cloud*
Guest Editors: B. B. Gupta, Dharma P. Agrawal, Nadia Nedjah, Gregorio Martinez Perez, and Deepak Gupta

✉ Bakhtgerey Sinchev
sinchev@mail.ru

[1] Almaty Technological University, Almaty, Kazakhstan

[2] The International Information Technology University, Almaty, Kazakhstan

[3] Almaty Management University, Almaty, Kazakhstan

[4] M Tynyshpaev Kazakh Academy of Transport and Communications, Almaty, Kazakhstan

[5] Abai Kazakh National Pedagogical University, Almaty, Kazakhstan

[6] Kazakh State Women's Teacher Training University, Almaty, Kazakhstan

formats is a challenging task. The number of customer requests, wrong orders and wrong deliveries may also increase [8]. Therefore, there is a need to implement syntactic and semantic constraints or defaults to avoid redundant statements or misinterpretations in big data.

In the landscape of database management systems, data analysis systems and transaction processing systems are separately managed, as they have different functionalities, characteristics and requirements [9]. The variation in data complexity gives rise to another problem– the collection and storage of large product data volumes. In order not to miss important information about the object, it is necessary to gather process and present big data in a form that is clear and easy to understand. Developing new methods for effective processing of big data is relevant to science and commerce, where data are integrated with business processes for user-centric management.

The purpose of this study was to develop a semantic approach towards big data processing. The approach in point is based on semantic methods selected by the help of mathematical statistics.

2 Background & Related Work

2.1 Big data challenges

In big data management, there are four challenges: volume, variety, velocity, and reliability [10]. Volume refers to the amount of data that needs to be processed. Variety covers different types of data, such as tabular data (databases), hierarchical data, documents, e-mail, metering data, video, images, audio, stock ticker data, financial transactions and more.

Velocity means how fast data is being produced and how fast it must be processed to meet the stakeholder demand. Reliability measures the accuracy and consistency of data. It is important because data sets come from different sources and thus may not fully meet the required standards of integrity [11]. The challenges of big data problems can be simplified with an ontological approach.

2.2 Semantic and statistical technologies

Nowadays, approaches that combine the semantic and statistical approaches to data processing are the most attractive ones [12]. These are hybrid computations that involve various algorithms that handle the same data: numerical-statistical (for example, deep learning) and logical-structural (including semantic ones) algorithms. The need for hybrid computations has already been acknowledged – when it comes to the operation based purely on statistical processing or logical conclusions, there will hardly be any progress in the issue, associated with artificial intelligence [13]. Thus, both approaches should be applied. However, there are only few examples of killer applications. Most likely, hybrid computations are used somewhere in text processing and language comprehension, when it is possible to combine the statistical machine learning and the precision of handling rules when processing certain crucial nuances related to the meaning. As an example, this is the polyglot persistence architecture, when the app handles both the schematic and non-schematic databases by using transactional and relational approaches in relation to a narrow range of problems [14]. Thus, the purpose of this research is to analyze and to develop an approach to mathematical modeling of ontology management.

2.3 Ontologies

Ontology as a declarative model of a certain problem domain is a central component of semantic-oriented intelligence. Problem domain complexity depends on the complexity of corresponding ontology. Thus, the known top-level ontologies reflect a significant number of concepts, for example, CYC – two million, and Wordnet – about 207 thousand. Complexity entails significant problems when it comes to ontology management problems. This problem class involves ontology creation, update, modification, visualization and validation, as well as origin documentation by components.

Ontology management problems lead to deterioration in the quality of ontology. In Bassaler et al. [15], ontology quality is assessed by the fulfillment of requirements regarding its completeness, correctness, and stability. Mistakes made by an expert during the complex ontology elaboration lead to nonrecognition of essential concepts and links in the problem domain entailing an incomplete and incorrect ontology creation.

Complex ontology management problems are studied in several directions. In particular, there are being developed metrics and methods to measure the ontology composition. In Hurwitz et al. [16], as with the software complexity, ontology complexity is defined as problems in performing such problems as ontology development, reuse and modification. Such work as Azarmi [17] is a meta-ontology called O2 – ontology as a semiotic object. Based on this ontology, three metrics for ontology complexity are developed: structural metrics, functional metrics and usability metrics. In Gandomi and Haider [18], there are introduced metrics for a pre-normalized ontology. Ontology normalization includes such steps as class (fact) naming, inheritance hierarchy materialization, name unification and attribute normalization. Such normalization has the purpose of converting various ontologies into a semantically equivalent form to create semantic complexity metrics.

Ontology visualization tools are being developed to increase the efficiency of any ontology management expert. They are based on the combinations of text, tabular, diagram and graph data mapping [19]. An important ontology management problem is to track the origin of ontology components and facts. Its solution is required if one has to validate and ensure the correctness of ontology, since the problem domain is changing. Thus, one has to track the dependencies between ontology components, facts available from the information base and corresponding domain objects. Currently, there are four levels of origin established [20]: static (constant data), dynamic (variable data), fuzzy (the origin of these data is by nature very fuzzy and unclear) and expert (expert analysis is required). The author of [21] puts forward an idea of tracking the origin of facts by recording the history of their changes and describing the events that had entailed them. Historically, problem ontologies were introduced as a result of problem analysis development. Problem analysis methods are used to determine and formalize all factors used by an expert solving a problem. Such methods are widely used to design computer program interfaces, in expert systems, and solution support systems [22]. In this case, the major purpose is to analyze and specify problem components, determine its structure and limitations.

Unlike other types of ontologies (general, domain ontologies), problem ontologies are created separately for similar problem classes, as well as the formalized concept of a related goal. Problem ontology research is closely related to conceptual modeling, as the formalized conceptual model of problem ontology is designed during problem ontology creation [23]. Both conceptual and ontological problem modeling have one important aspect – the interaction with a domain expert that creates and validates the ontology. Ontology research has involved modeling environments that allow creating and implementing ontological models for individual problems. Currently, the major research in the field of ontological modeling is devoted to declarative ontologies – general, domain ontologies [24]. Problem ontology direction is not sufficiently developed. On the other hand, existing research in the field of problem ontology considers ontology creation for individual problem classes. At this approach, there are restrictions on the ontology transferability and reuse to solve the problems in other problem domains, since the same entities will be interpreted differently by different problem ontologies. We will refer to problem ontologies, based on a particular general ontology, as to ontological models in order to show this dependence and avoid ambiguities. Ontological models made it possible to simplify the solution of complex ontology management problems. The purpose of this research is to find ways for ontological models to simplify the complex ontology management and improve the ontology quality.

3 Big data modelling techniques

This section introduces a conceptual mathematical framework for ETL (Extract-transform-load) processes that is built upon semantic technologies.

3.1 Mathematical representation

Here is the one of the possible ways of ETL formalization with regard to applied mathematics. Initially, let us consider the widely used type of functional dependencies:

$$y\,(t) = f\,(x(t), t), \tag{1}$$

Where: *x, y, f* - vectors with *nx1, mx1, mx1* dimensions, *f* - known vector function, *t* -time. The *x* variable serves as a recoverable input time-dependent data, *f* serves as the transformation process function, *y* – load output data.

Next, let us consider the system of differential equations.

$$\dot{x} = f\,(x(t), u(x,t), t) \tag{2}$$

Where: initial condition $x\ (t_0) = x_0$ is correct at the time interval $[t_0,\ t_1]$; $x(t_0)$ – input data˙, *x* – the output data. In this system, *u(x,t)* is controlled over by a computer or by the control unit, included into the original system.

The system equilibrium can be described by linear and nonlinear algebraic equations

$$f\,(x(t), t) = 0. \tag{3}$$

In other words, they are the operating modes of the controlled objects (system). Random functional series can be expressed as:

$$y(t) = \sum_{i=0}^{\infty} x_i(t) \tag{4}$$

It is known that many continuous functions are described by this functional series. For example, the sine/cosine is expressed through the harmonic (power) series. In turn, components of the power series can be found by the interpolation formulas, introduced by Lagrange, Newton etc.

Thus, both input and output data can be calculated for a given moment of time and there is no need to store them digitally.

The basic operation intended for the text data implies the extraction of useful information without converting the original data by the index terms, template or mask.

3.2 Information modelling

Any computer system transforms (converts) the information. Such a system has an input through which it receives information to be processed, and an output, which provides the output information generated by the computer system in response to the relevant input information.

In a functional sense, human intelligence works similar to a digital computing system. Both systems work with a finite set of multi-dimensional information.

Information modelling can be carried out by means of classic alphabetical operators, when the following two features are not important: 1) their infinite domain; 2) input and output languages of a classic alphabetical operator may only contain the words with equal length. If we introduce the finite dimensionality into the definition of an alphabetical operator, we will get the concept of a finite alphabetical operator. At this point, input and output languages could include words of different length, thereby complicating the mathematical language for recording such operators.

Formal description of the natural and artificial intelligence systems requires such mathematical tools that would provide a convenient record for any final alphabetical operator. Based on these considerations, there has been developed the algebra of finite predicates [25]. The definition of a final predicate implies the following [25]:

Let us assume that A is a finite alphabet, which contains k set of letters a_1, a_2. …a_n, $\sum$− − set consisting of two components designated by symbols *0* and *1*, and called false and true, respectively. The variable over the A set is a literal variable, while the variable over the $\sum$ set – a logical variable. The finite-local predicate over the alphabet A is represented by any function $f\ (x_1,\ x_2,\ \ldots,\ x_n) = t$ with n letter arguments $x_1,x_2,\ldots,x_n$, over the A set, which takes logical values t.

As can be seen from this definition, values of finite predicate variables, unlike the values of variables related to the finite alphabetic operators, presented by words, are literal. The switch to the alphabetic variables provides the possibility to develop a convenient mathematical language to describe various intelligent systems.

3.3 Data mining

The basic concept of text data mining methods centers around the similarity of objects, as well as around on the quantitative measure. The key techniques involve the following.

The first method is term-based. It is effective in computational performance and involves seeking a word in a document that has semantic meaning. This technique, however, has disadvantages such as polysemy and synonymy [36], where polysemy is a single word having multiple meanings and synonymy refers to multiple words having the same meaning. The next popular method is the phrase-based technique. Since a phrase has more meanings and is less ambiguous, this method performs better than the former one. However, it also has disadvantages such as inferior statistical properties to terms, low frequency of occurrence and the presence of excess information that is not related to requests. The more complicated and constructive methods are concept-based and pattern taxonomy methods. The first is based on sentence- and document-level analysis [37]. It stands on the following three components: semantic analysis of a sentence, building of a conceptual ontological graph, and concept extraction. The concept-based method allows differentiating between important and unimportant words, which are widely used to process the natural language. The pattern taxonomy method involves patterns with the 'is-a' relation between them [38]. It can be effective and accurate if patterns such as image signals are correctly selected. Signal image is a set of primary signs – the results of direct measurements or observations. The signal image or the dependent secondary characteristics are the initial data used to take one of the possible decisions regarding the object, for example, regarding its belonging to one of the specified classes. There are logical recognition methods, which imply information processing according to a well-defined algorithm to emphasize valuable information and intuitive recognition methods when valuable information is generated.

Semantic analysis plays an important role in the logical recognition methods [26], as a set of operations that support the comprehension of a natural sign system (pictures, phenomena or texts), introduced as a record, by using some kind of a formalized semantic language. This approach provides the possibility to define a new problem, which implies studying the impact of external factors on the price strategy of the enterprise. Its structure includes two sub-problems: processes of determining factors (market events) and obtaining association rules in a specified sector within a specified time limit. Association rules describe the relationship of factors occurring in a specified segment at a certain moment or period.

The first problem can be formulated as building a syntactic model of the Internet news analysis and identifying a unique market event by clustering, based on metric proximity of the two news blocks.

The second sub-problem implies obtaining association rules. This new approach is based on the idea that online news can be viewed as a marketing data container, which includes various external factors. Based on these factors, fitted on the traditionally collected internal data of the enterprise, one can create a set of rules that would specify, for example, the predicted values of the indicators. In this case, thus, the first problem is to identify market events that are significant for decision-making.

Thus, based on the morphological and syntactic analysis [27], we have formulated an original approach to identifying such market events. This approach is applied through a plurality of syntactic patterns, obtained by the domain ontology. These models take into account the category of market events (external factors): consumption and demand, competitor's profile, inflation, international prices, R&D, consumer profile, consumer psychology etc.

4 The process of model development

The ontology is a mandatory step and allows generating a plurality of semantic fields, syntactic patterns and tokens in accordance with the subject area (certain specified market).

For example, the ontology of events that have occurred in the raw material market with elastic demand (Fig. 1) allows building a syntactic model that describes the competitor's profile category. The model includes a set of syntactic patterns made for phrases, divided into verbal and noun phrases, as well as many tokens, based on morphological analysis.

The phrasal text elements are handled with regard to their grammar. The initial elements can be allocated within a sentence. Suffixes or tokens are the center units in the analysis. In order to extract data from all news flows, similar models (grammar rules) should be formed for each news category.

We suggest a two-step processing of online news for identifying random external factors. Firstly, this implies classification, made through the syntactic and morphological analysis with regard to the *M, E, G* sets. As a result, we get the values for the event category vector $\overrightarrow{c}$. Each element c_k, $k = 1, K$ of this vector will take a value of 1 or 0, if the news refer to the *k*-category or another category, respectively.

The second phase implies the allocation of similar event clusters that would allow avoiding duplication and story chains, thus obtaining a stream of high-quality events. One can determine whether two news-duplicates are unified in one event by tracking the coincidence of coordinates c' of one news and with the coordinates c" of another news. Their release dates should vary within the threshold value d_t that is sufficient for reflecting market dynamics. In other words, inequation $|d' - d''| \leq d_t$ should be valid, where *d'* and *d''* are dates of the first and the second news releases. The proximity assessment of two news by tokens, taken from M, E and G, is carried out by the following formula:

$$F_p = \left[1 + \sum_{i=1}^{I}\left(m_i{'} - m_i{''}\right)^2\right]\left[1 + \sum_{j=1}^{J}\left(e_j{'} - e_j{''}\right)^2\right]\left[1 + \sum_{h=1}^{H}\left(g_h{'} - g_h{''}\right)^2\right] \tag{5}$$

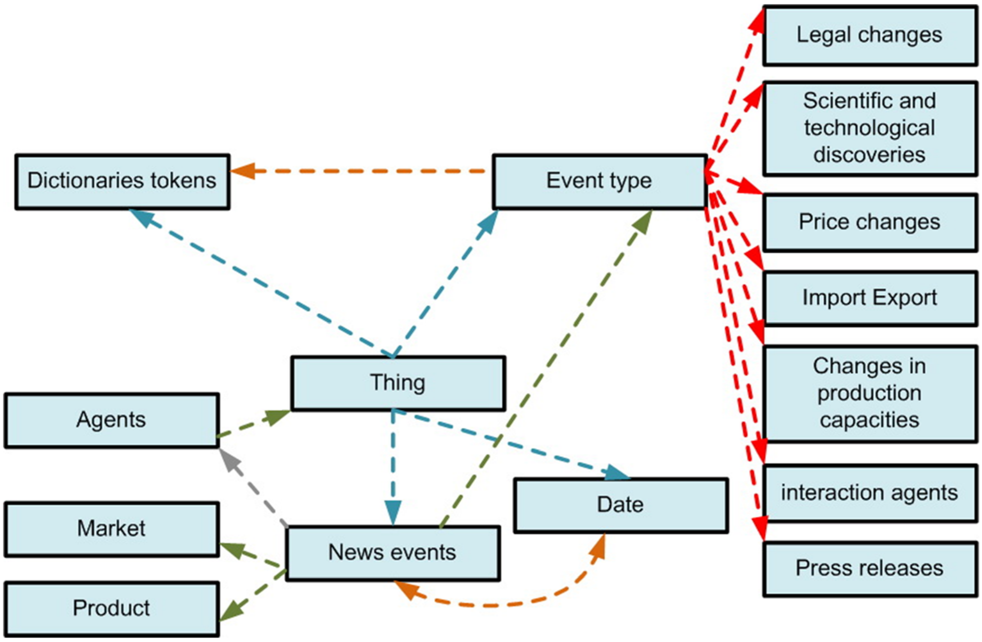
Where: $m_i{'}, m_i{''}$- coordinates of the vectors $\overrightarrow{m'}$ and $\overrightarrow{m''}$ that were formed in relation to the linearly ordered token sets $\widetilde{M} = M^1 \cup M^2$ of both releases with similar dimensionality $l = |\widetilde{M}|$; M^1 and M^2 are unordered token sets of each release. As for the first news, coordinates of $\overrightarrow{m'}$ take the following value:

$$m_i = \left\{\begin{matrix} 1, l_i^M \in M^1, \\ 1, l_i^M \in M^1 \end{matrix}\right\}, \tag{6}$$

Where: l_i^M - a token, taken from *M,* and referring to all possible market-related tokens formed during the study of news domain. In this regard, $|M| > l$.

The vector $\overrightarrow{m''}$ for the second news release is formed in a similar way. The coordinate vector $\overrightarrow{e'}$ with relevant coordinates $e_j{'}$ and $e_j{''}$ ї is formed in relation to the tokens l_i^E taken from the *E* set, containing all possible tokens of counter-agents pursuant to the linearly ordered set, $\widetilde{E} = E^1 \cup E^1$, where E^1 and E^2 are unordered sets of counter-agent tokens with similar dimensionality $j = |E|$. In this regard, $|E| > j$. The news market geography vectors $\overrightarrow{g'}, \overrightarrow{g''}$ are formed with the relevant coordinates $g_h{'}$ and $g_h{''}$ in relation to tokens l_h^G taken from the *G* set, containing all possible event geography tokens pursuant to the linearly ordered set $\widetilde{G} = G^1 \cup G^2$, where G^1 and

Fig. 1 News ontology fragment



G^2 are unordered sets of event geography tokens with similar dimensionality$H = |G|$. In this regard, $|\ G\ | \ > H$.

In most cases, token sets N_1, related to the first news release $N_1^1 = M^1 \cup G^1 \cup E^1$ and the second one $N_1^2 = M^2 \cup G^2 \cup E^2$, will be different. Therefore, there will be no full coincidence (when the formula (1) equals to 1). This problem can be solved either by an expert assessment of news unification thresholds or by calculating a limit value through analysis.

We suggest making an analytical calculation of permissible error, as this will allow assessing the news similarity by the initial predicate. Such calculation is based on a set of additional tokens. The set refers to tokens, which are included into the news, but do not cover the event initially; they only specify its features, in particular, indicate its time-related feature, character, impact etc. Further calculation implies the introduction of intermediate assessment proximity degree by:

$$F_\alpha = \left\{ \begin{matrix} |N_3^1 \cup N_3^2 \backslash N_3^1 \cap N_3^2 \| N_3^1 \cup N_3^2 \backslash N_3^1 \cap N_3^2| = 0; \\ 1, |N_3^1 \cup N_3^2 \backslash N_3^1 \cap N_3^2| = 0, \end{matrix} \right\}, \tag{7}$$

Where: are N_3 sets of the first and the second news release. The coefficient can be get from (7):

$$\alpha = |N_3^1 \cup N_3^2| / F_\alpha. \tag{8}$$

The coefficient indicates the news proximity by irrelevant tokens: α increases as the news proximity does. Semantic meaning of the coefficient (8) implies the fact that its increase points to the probability that the released news describe the same event. This probability arises in the light of links between the words in natural languages. Therefore, first and the second predicates impose less requirements imposed for proximity.

The predicate $F_\alpha \leq \alpha$ makes it impossible to compensate the divergence by F_α through the growth of α, as the growth rate of is much higher than α; besides that, α is a finite quantity and takes a value $\alpha = \left[1, |N_3^1 \cup N_3^2|\right]$. Unlike α, higher growth rate of F_p indicates the higher specific gravity of F_p over . Therefore, N_1 tokens have higher specific gravity, than the tokens.

Table 1 shows the behavior of $F_p \leq \alpha$ \components for the case when =[1, 10], where n_p is the number of differences in word tokens, namely – $|N_1^1 \cup N_2^2 \backslash N_1^1 \cap N_2^2|$, and n_α is the number of coincidences, namely – $|N_3^1 \cap N_3^2|$.

The precise assessment of news proximity degree requires the introduction of N_2 sets with restored tokens. The sets N_1^1 and N_2^2 (first and the second news releases, respectively) are formed with regard to and, but with new collected domain information added. For example, one can consider adding data on products and geographic markets of agents. Thus, the secondary condition for assessing the news proximity degree was introduced upon the reconstructed token vectors, namely, $F_s \leq \alpha F_p$; its left part is calculated according to the formula, which is similar to (1), but based on the sets N_2^1 and:

Table 1 The values of and α

n_p	0	1	2	3	4	n_α	0	3	5	7	8	9	10
F_p	1	2	4	6	12	α	1	1.2	1.5	2.1	3	5.5	10

$$F_s = \left[1 + \sum_{i=1}^{I} \left(m_i{'} - m_i{''}\right)^2\right] \left[1 + \sum_{j=1}^{J} \left(e_j{'} - e_j{''}\right)^2\right] \left[1 + \sum_{h=1}^{H} \left(g_h{'} - g_h{''}\right)^2\right] \tag{9}$$

The formulas (5), (8), (9) are united into a single complex formula, based on the prerequisites related to the coincidence of event categories and news proximity within d_r

$$F = \left\{ \begin{matrix} \overrightarrow{c} = \overrightarrow{c}'; \\ |d' = d''| \leq d_t; \\ F_p \leq \alpha; \\ F_s \leq \alpha F_p \end{matrix} \right\} \tag{10}$$

The predicate (10) is interpreted in the following way: $F_p \leq \alpha$ means that news proximity should be evaluated at first by the N_1 token set and by other token sets in case when the first evaluation has indicated high proximity. Indicates that difference between proximity estimates, obtained by the N_1 and N_2 token sets, should be within the error α limit.

Although the growth rate of F_s, F_p is similar, inequation $F_s \leq \alpha F_p$ indicates a higher specific gravity of F_p over F_s, namely – a higher specific gravity of token sets N_1, not N_2 – by combining two news in a cluster.

The news flow, obtained after classification and clustering, will have almost one-to-one correspondence with the real events that gave rise to relevant news.

In order to forecast, one has to develop a set of association rules for a received flow of events. At this point, let us introduce an additional notation. Let us consider that a market event happened during the time segment $\tau - Y_i^\tau$, $i = 1, 2$. This event caused changes, for example, in price. We denote the additional event Y_0^τ, which reflects the line of change:

$$Y_0^\tau = \left\{ \begin{matrix} +1, p^{\tau+1} - p^{\tau-1} > \widetilde{p}_m; \\ -1, p^{\tau-1} - p^{\tau+1} > \widetilde{p}_m; \\ 0, |p^{\tau+1} - p^{\tau-1}| \leq \widetilde{p}_m; \end{matrix} \right\} \tag{11}$$

Where: $\widetilde{p}_m$ – minimum fluctuation threshold of projected indicator for a specific market, $\widetilde{p}_m > 0$.

In this case, the problem of change forecasting is interpreted as a problem of finding the sequence of specific market events:

$$Y_i^\tau Y_j^\tau \rightarrow Y_k^\tau \rightarrow Y_0^\tau. \tag{12}$$

The sequence (12) is called a rule. The rule (12) shows that after the simultaneous occurrence of events Y_i^τ and Y_j^τ, event Y_k^τ, leads to the event Y_0^τ according to the formula (11), where $i, j, k \in Z$. The rules of this type (12) can be built by means of the SPADE algorithm. Then, one can obtain the final value of a projected indicator according to association rules, based on the online news analysis that is carried out through the identification of current market conditions and rules relevant for this (current) situation.

5 Results and discussion

Thus, the problems of big data processing can be solved by an approach, based on the problem decomposition. We suggest allocating two problems: status identification and search for association rules. They were solved to illustrate how the market analysis is done by processing the array of online news, related to a specific topic. The prospect of this approach implies the following. Firstly, status identification remains urgent regardless of the problem domain. This problem can be solved by using the artificial intelligence tools. Secondly, search for regularities in large arrays of accumulated data allows collecting additional information for decision-making.

Let us denote the set of such rules like (12), as $\in R$, where τ is a sequence (or set) of market events that occur before a change in the controlled parameter (price). Based on these, allocated factors may somehow affect the final value of a projected indicator within a specified market segment. In this case, we can identify the market situation leading to a predetermined value as a corresponding rule τ. In this context, there are two attributes corresponding to each obtained rule: *s* - supportability, characterizing the absolute frequency rule in the original sample; *c* – accuracy, namely – the risk of changes in the value on the background of an emerging set of events, described by the rule .

Supportability and accuracy are two important measures with the following definitions. Supportability is a number or percentage of transactions, containing a specific set of data. It reflects the frequency of element combinations. Practical problems, especially those related to customer data processing, are solved through the identification of a minimum support for the association rules. Thus, the set is of interest, if its supportability exceeds the user-defined minimum. The accuracy of association rules defines the probability of a chain of certain events to occur. The rule accuracy reflects the percentage of transactions that keep an object in the set.

However, this approach brings up an uncertainty problem at the stage of identifying the current situation and choosing rules, since association rules τ are characterized by different degree of accuracy and supportability. Any rule can have a very high supportability (obvious rule), and, in contrast, very low supportability (non-obvious rule). Consequently, the forecast quality depends on the identification method.

The introduced technology has been applied in the context of price strategy development. The sample size of price values has amounted to 800 items; the sample size of Internet news – 2700 items, while the first 600 price values and the corresponding 2100 news were taken as a training set. The two test samples were formed from the remaining values. The experiments allowed evaluating the effectiveness of an introduced technology. The methods of high-quality forecasting were assessed upon the built models with minimized value of the likelihood function. Many association rules were obtained through the SPADE algorithm. In order to assess the accuracy of introduced forecasting method, we have considered the relevant methods (Table 2).

Data analysis shows that forecasts, made by using association rules, are by 6% more accurate than the forecasts, made by using pattern conventional methods. Greater accuracy is achieved because the forecasts, based on association rules, allow considering those events that affect prices in the predicted value, while the regression methods contribute to an indirect consideration.

A number of related works showed that semantic technologies are effective in big data processing. It was shown that eClass ontologies could be integrated if special dictionaries were used [28]. Some authors showed how a big data management system can be improved with a semantic web technology [29]. This technology, however, did not allow an efficient and scalable data system. A semantic approach was applied to create the architecture of the Aletheia system, which enabled the integration of structured and unstructured information [30]. A semantic model ensures data exchange and conversion through an Internet service hub. Using an integration-oriented approach and GoodRelations ontologies [31] makes it possible to convert data from the BMEcat format. This allows improving the quality of data and restoring some information. These methods are also in good agreement with the multistage mining approach applied in this study.

The systematic analysis was applied to various tools necessary for integrating data from different sources [32, 33]. It was found that the most effective solution is to use similarity tools, which can be aimed at matching terms, graphs, etc.

Table 2 Experimental error values

	Sample	
Method	Control 1	Control 2
Exponential smoothing	38.2%	39.5%
ARIMA	37.4%	37.1%
Association rules	30.9%	31.2%

These technologies allow an automatic control over big data processing and integration. The relationship between big data trends and environmental issues was addressed in another review paper [35], which threw light onto an innovative approach to big data management in the field of energy-efficient green technologies. The approach in point may provide promising results. Unlike the association rule-based method used in this study, the above approach may be less effective if applied to complex systems.

6 Conclusions

Research results indicate that the future of data science and Big Data Analytics is in the latest achievements made in applied mathematics and applied in data processing modelling, regardless of data volume. In other words, semantic methods and mathematical statistics and vectorization, if applied together under a fact-oriented approach to data, produce information of practical value. The introduced approach is successful because of well-developed methods of mathematical statistics. The representation changing technique is useful for rapid data processing. However, they can be applied to non-homogeneous and unstructured data under a semantic approach, applied to generate information that is suitable for handling the nonhomogeneous data. The fact-oriented approach and methods of hybrid data processing are at the starting point of their history. The main data mining models, considered in this article, show that this problem can be solved in several stages, most important whereof include data identification and search for regularities (rules). The text data, such as online news, can be identified by means of semantic models that contribute to the text proximity assessment not only by the coincidence of certain words, but also by their semantics. Therefore, duplicates are avoided and data are clear for further processing. In this research, the second stage of data processing implies building association rules that would allow identifying the chain of events that significantly affect the analyzed event. Based on the market prices analysis and forecast, this research shows that the original approach allows improving the forecast accuracy by 6%. we hope that our approach will be useful in this field.

Availability of data and material Data will be available on request.

Funding information This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Compliance with ethical standards

Conflict of interests Authors declare that they have no conflict of interests.

References

1. Chen CP, Zhang CY (2014) Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf Sci 275:314–347
2. Laudon KC, Laudon JP (2015) Management information systems. Upper Saddle River, Pearson
3. Zaurbekov N, Aidosov A, Zaurbekova N, Aidosov G, Zaurbekova G, Zaurbekov I (2018) Emission spread from mass and energy exchange in the atmospheric surface layer: two-dimensional simulation. Energ Source Part A 40(23):2832–2841
4. Kwon O, Lee N, Shin B (2014) Data quality management, data usage experience and acquisition intention of big data analytics. Int J Inf Manag 34(3):387–394
5. Bulat PV, Zasuhin ON, Uskov VN (2012) On classification of flow regimes in a channel with sudden expansion. Thermophys Aeromech 19(2):233–246
6. Deng Q, Gönül S, Kabak Y, Gessa N, Glachs D, Gigante-Valencia F, Thoben KD (2019) An ontology framework for multisided platform interoperability. In: Popplewell K, Thoben KD, Knothe T, Poler R (eds) Enterprise interoperability VIII. Proceedings of the I-ESA conferences, vol 9. Springer, Cham
7. Rocha, V, Varela, L, Carmo-Silva, S (2016). Sharing product information for supporting collaborative product development. Dept. Production and Systems, School of Engineering, University of Minho, Braga, Portugal
8. Cunha FA, dos Passos Silva J, de Barros AC, Romeiro Filho E (2017) The use of information management tools as support to the product development process in a metal mechanical company. Product: Manag Develop 11(1):33–41
9. Welzer, T, Eder, J, Podgorelec, V, Latifić, AK (2019). Advances in Databases and Information Systems. In: 23rd European Conference, ADBIS 2019, Bled, Slovenia, Vol. 11695. Springer Nature
10. Beyer, M (2011). Gartner says solving "big data" challenge involves more than just managing volumes of data
11. Zikopoulos, PC, deRoos, D, Parasuraman, K, Deutsch, T, Corrigan, D, Giles, J, Melnyk, RB (2011). Harness the power of big data—The IBM Big Data Platform. McGraw-Hill
12. Wu X, Zhu X, Wu GQ, Ding W (2013) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107
13. Abacha, AB, Zweigenbaum, P (2011). Medical entity recognition: A comparison of semantic and statistical methods. In: Proceedings of BioNLP 2011 Workshop, pp. 56–64. Association for Computational Linguistics
14. Wiese, L (2015). Polyglot database architectures= Polyglot Challenges. In LWA, pp. 422–426
15. Bassaler, J, Zaïm, S, Prémont, C (2014). What can businesses do to capture the full potential of big data? Orange business services
16. Hurwitz, J, Nugent, A, Halper, F, Kaufman, M (2013). Big Data for Dummies. Wiley
17. Azarmi, B (2016). Scalable big data architecture. Apress
18. Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manag 35(2):137–144
19. Tian X, Han R, Wang L, Lu G, Zhan J (2015) Latency critical big data computing in finance. Journal of Finance and Data Science 1(1):33–41
20. Sukhobokov AA, Lakhvich DS (2015) The impact of big data tools on the development of scientific disciplines related to modeling, science and education. Online journal of N.E. Bauman MSTU 3: 207–240
21. Barlow, M (2013). Real-time big data analytics: emerging architecture. O'Reilly
22. Thaduri A, Galar D, Kumar U (2015) Railway assets: a potential domain for big data analytics. Procedia Comput Sci 53:457–467

23. Karimi, HA (2014). Big data: techniques and Technologies in Geoinformatics. RC Press
24. Klemenkov PA, Kuznetsov SD (2012) Big data: current approaches to storage and processing. Proceedings of the Institute for System Programming of the Russian Academy of Sciences 23:143–156
25. Hutter M (2005) Universal Artificial Intelligence. Springer, Berlin
26. Evangelopoulos NE (2013) Latent semantic analysis. Wiley Interdiscip Rev Cogn Sci 4(6):683–692
27. Seeker W, Kuhn J (2013) Morphological and syntactic case in statistical dependency parsing. Comput Linguist 39(1):23–55
28. Hladik, J, Christl, C, Haferkorn, F, Graube, M (2013). Improving industrial collaboration with linked data, OWL. In: OWLED
29. Brunetti JM, García R, Auer S (2013) From overview to facets and pivoting for interactive exploration of semantic web data. IJSWIS 9(1):1–20
30. Wauer, M, Schuster, D, Meinecke, J (2010). Aletheia: an architecture for semantic federation of product information from structured and unstructured sources. In: Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services, pp. 325–332
31. Stolz, A, Rodriguez-Castro, B, Hepp, M (2013). Using BMEcat catalogs as a lever for product master data on the semantic web. In: Extended Semantic Web Conference, pp. 623–638. Springer, Berlin, Heidelberg
32. Otero-Cerdeira L, Rodríguez-Martínez FJ, Gómez-Rodríguez A (2015) Ontology matching: a literature review. Expert Syst Appl 42(2):949–971
33. Dragisic Z, Ivanova V, Li H, Lambrix P (2017) Experiences from the anatomy track in the ontology alignment evaluation initiative. J Biomed Semant 8(1):56
34. Wu J, Guo S, Huang H, Liu W, Xiang Y (2018) Information and communications technologies for sustainable development goals: state-of-the-art, needs and perspectives. IEEE Commun Surv Tut 20(3):2389–2406
35. Wu J, Guo S, Li J, Zeng D (2016) Big data meet green challenges: Big data toward green applications. IEEE Syst, J. 10(3):888–900
36. Singhal, A, Buckley, C, Mitra, M (2017). Pivoted document length normalization. In: Acm sigir forum, pp. 176–184. New York, NY, USA, ACM
37. Shehata, S, Karray, F, Kamel, M (2006). Enhancing text clustering using concept-based mining model. In: Sixth International Conference on Data Mining (ICDM'06), pp. 1043–1048. IEEE
38. Wu, ST, Li, Y, Xu, Y, Pham, B, Chen, P (2004). Automatic Pattern-Taxonomy Extraction for Web Mining. In: IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242–248

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.