УДК 004.942
DOI: 10.12737/article_5a3779fddad627.59611142

Я.А. Швецов, В.И. Аверченков, М.Ю. Рытов, В.П. Фёдоров, Г.А. Федяева

ИНТЕЛЛЕКТУАЛЬНАЯ ОБРАБОТКА ДАННЫХ В ЗАДАЧЕ СИСТЕМАТИЗАЦИИ ЭКОНОМИКО-СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ

Рассмотрены основные принципы интеллектуальной обработки больших массивов данных. Предложены структурная и функциональная схемы программного комплекса, осуществляющего работу со статистической информацией, полученной с сервера фондовой биржи и новостных сайтов. Представлены математические модели для обработки текстовых данных, а также проведения корреляционного анализа динамики биржевых временных рядов. Подробно описаны подходы для передачи и обработки значительных объёмов экономической информации при помощи автоматизированных интеллектуальных систем.

Ключевые слова: интеллектуальная обработка текста, корреляционный анализ, коэффициент корреляции, информационные системы, фондовый рынок, котировки акций.

Ya. A. Shvetsov, V.I. Averchenkov, M.Yu. Rytov, V.P. Fyodorov, G.A. Fedyaeva

INTELLIGENT DATA PROCESSING IN PROBLEM OF ECONOMIC STATISTICAL INFORMATION SYSTEMATIZATION

The purpose of this paper is a description of fundamental moments in the course of obtaining and processing economic statistical information. Besides, in this paper there is offered an information system allowing the negotiation of basic problems arising during the work with large data files on stock quotations in a stock market. Basic principles of an intelligent approach to large data file processing are described, the analysis of a current state of a subject field is given. On the basis of the investigation the conclusions on a topicality of the problem under consideration have been drawn, and basic approaches to its solution are defined. For the work within the limits of problems set there was widely used a methodology and theory of intelligent data processing both a text format, and a numerical one. In addition to the consideration of theoretical aspects in the functioning of a software complex meeting parameters specified in the paper the process of designing an information system for the carrying out a thorough analysis of statistical information is manifested. The paper reports the structural scheme of a complex and also a flow block describing thoroughly functions of the developed automated system. The efficiency of used methodological and mathematical principles of information system functioning for statistic data processing is substantiated. At the end of the paper there is drawn a conclusion on a correspondence of methods used to those processes within the limits of which a software system has to function.

Key words: intelligent txt processing, correlation analysis, correlation coefficient, information systems, stock market, stock quotations.

Многообразие функциональных возможностей современных информационных технологий обусловливает их активное использование для обработки и анализа значительных объёмов информации. Особенно остро стоит проблема создания эффективных моделей и методов для автоматизации процесса поиска неочевидных, объективных и практически значимых закономерностей в больших информационных массивах.

Традиционные подходы оперативной аналитической обработки данных (OLAP) в основной своей массе ориентированы на проверку заранее сформулированных гипотез и проведение разведочного анализа методом простого перебора наиболее вероятных вариантов. Однако помимо них в последние годы получили широкое распространение методы интеллектуального анализа данных, основной задачей которых является поиск неочевидных закономерностей в массиве слабоструктурированных данных. Инструменты такого подхода позволяют строить гипотезы о взаимосвязях внутри сложных объектов на основании заложенного в информационную систему алгоритма [1]. Преимущества интеллектуального анализа по сравнению с другими методами обработки данных очевидны. Большинство статистических методов для выявления взаимосвязей в данных используют концепцию усреднения по выборке, приводящую к операциям над не-

существующими величинами, тогда как методы интеллектуального анализа оперируют реальными значениями. Указанный подход позволяет, опираясь на ретроспективные данные, прогнозировать будущие состояния целых групп объектов.

В основу рассматриваемой технологии положена концепция шаблонов, представляющих из себя закономерности, свойственные подвыборкам данных и выраженные в форме, понятной человеку. Интеллектуальные методы обработки статистической информации образуют несколько тесно взаимосвязанных разделов:

- Подготовительный анализ, направленный на установление природы и структуры массива данных.
- Выявление основных закономерностей и внутренних связей объектов (корреляционный и регрессионный анализ).
- Динамические модели и прогнозирование изменения состояния объектов на основе показателей временных рядов.
- Многомерный статистический анализ (кластерный, компонентный и факторный анализ).

Каждый из представленных видов интеллектуальной обработки информации позволяет выявить неоднозначные гипотезы в выборке данных, однако лишь комплексное использование нескольких подходов одновременно позволяет получить более полную картину происходящего [2]. В качестве примера задачи по нахождению закономерностей в слабоструктурированном массиве данных можно выделить исследование котировок акций, валют и товаров на фондовом рынке. В обычном виде статистическая информации представляет из себя набор значений изменения цены за определённый интервал. Построение и исследование свойств временных рядов, нахождение корреляции между показателями различных эмитентов и использование математических моделей для описания поведения целых экономических отраслей, в свою очередь, позволяют выдавать информацию в понятном человеку формате [3].

На рис. 1 представлена структурная схема предлагаемого программного комплекса под рабочим названием «Фондовая аналитика». На ней наглядно показана взаимосвязь основных элементов системы, таких как база данных, а также модули получения и обработки биржевой информации. Кроме математических моделей программа проводит семантический поиск по новостным серверам, позволяя пользователю получить наиболее полную картину происходящего на фондовом рынке. Данные представляются в виде графиков и диаграмм, демонстрируя основные тенденции изменения котировок акций, товаров и валют.

В основе предлагаемой методики моделирования поведения экономических субъектов лежит так называемый технический анализ котировок. Его инструменты позволяют прогнозировать изменение стоимости акций и товаров в будущем при помощи анализа предыдущих значений ценовых характеристик. Технический анализ основан на анализе временных рядов, чаще всего графиков с различным значением расчётного интервала. Кроме того, во внимание принимается информация об объемах торгов и другие важные статистические данные. В техническом анализе применяются разнообразные инструменты и методы, но все они основаны на одном общем предположении: анализируя временные ряды посредством выделения трендов, возможно спрогнозировать поведение цен в будущем [4].

Технический анализ хорошо себя проявляет на сильно волатильных рынках, поэтому чаще всего его применяют при анализе товарных и финансовых рынков. В частности, большие объемы обращаются на следующих финансовых инструментах: нефть, золото, природный газ, серебро, евро, английский фунт стерлингов. При анализе товарных рынков хорошо показывает себя совмещение технического и фундаментального анализа. Так, например, пшеницу выгоднее всего покупать осенью - после преодоления традиционного летнего тренда на снижение цен. После сбора мирового урожая, в сентябре, цена на сельхозпродукцию находится в районе минимальных цен, а конкретный момент покупки нужного объёма продукции можно обнаружить путём выявления тренда.

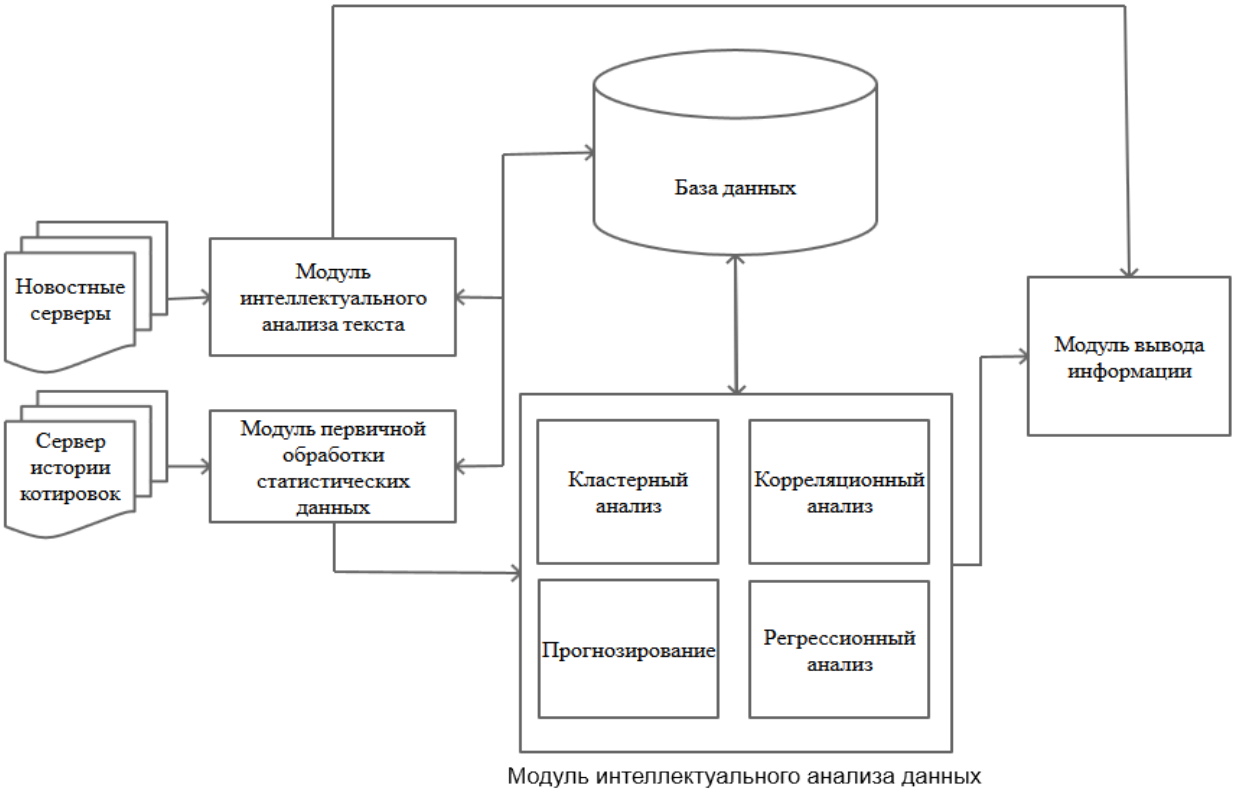

Рис. 1. Структурная схема программы для интеллектуальной обработки котировок «Фондовая аналитика»

Именно агрегирование полученной информации в удобную для восприятия форму и является рассматриваемой в текущем исследовании задачей. Проведение интеллектуального анализа всего объёма предоставленной в общем доступе информации позволяет улучшить качество процесса поддержки принятия управленческих решений в ряде смежных областей.

Методы проведения кластерного анализа на заданном массиве данных, а также принципы регрессионного и кластерного анализа данных о котировках фондового рынка подробно рассмотрены в других работах [5;6], поэтому подробно остановимся на остальных модулях системы, таких как модуль интеллектуального анализа текста и модуль корреляционного анализа временных рядов.

Методы интеллектуального анализа текста объединены под общим названием «Текстовая добыча» (Text Mining). Основная задача использования этой технологии заключается в предоставлении моделей обработки неупорядоченной информации для выделения из неё значимых числовых показателей и формирования данных для последующего анализа посредством математических методов.

Существует целая группа подходов к увеличению производительности подобного анализа путём предварительной обработки входной выборки данных. Представим некоторые из них.

Создание базы данных терминов. Индексирование загруженных из сети Интернет текстов занимает продолжительное время, поэтому технология интеллектуальной добычи информации предполагает создание базы данных, содержащей основные индексы слов, выражений и даже целых документов [7]. Это позволяет создать автоматизированную систему, в которой информация, полученная из обучающего множества текстов, используется для создания математических моделей и прогнозирования будущих состояний объектов за пределами обучающей выборки.

Ограничение на поиск некоторой последовательности символов. Перед началом обработки исходных текстов необходимо задать некоторое количество параметров, определяющих основные характеристики входного массива данных. В

общем виде это означает наложение ограничений на работу автоматизированной системы с определённой последовательностью символов. Так, например, можно исключить из анализируемого множества слова, начинающиеся с заглавной буквы, слишком длинные семантические единицы или даже отдельные числа и символы. Кроме того, существуют подходы, при которых устанавливается определённая планка, фиксированный предел для ограничения минимального процента появления определённого слова или фраз в тексте. Также можно исключить слова, которые короче или длиннее фиксированного предела, или часто используемые слова, не несущие в себе значимой информации (предлоги, союзы, вводные слова).

Объединение близких по значению слов и словосочетаний. Анализ текста более точен при объединении синонимов или даже целых фраз со сходным значением в одну семантическую единицу [1].

Алгоритмы морфологического анализа. Одним из важнейших этапов интеллектуального анализа текстов является выделение однокоренных слов в отдельную группу и дальнейшее рассмотрение такой группы в качестве общей семантической единицы.

Мультиязычность. Распознавание иностранных слов и заимствований необходимо для более точного анализа загруженного в автоматизированную систему текста.

После индексации входного массива текстовых данных и определения частоты использования семантических единиц применяются дополнительные математические методы для получения агрегированной информации.

Частота появления семантической единицы обычно отражает важность этой группы слов в исследуемом тексте [7]. Однако нельзя предполагать, что сами индексы частоты слов пропорциональны важности соответствующего слова. Поэтому в ходе интеллектуального анализа текстов для каждой семантической единицы вычисляют преобразованную частоту (wf):

$$f(wf) = 1 + log\ (wf)$$ (для wf > 0).

Это преобразование уменьшает абсолютные значения исходной частоты появления слов в тексте, а следовательно, и их влияние на последовательные вычисления [8].

Кроме того, при интеллектуальном анализе текстов активно используется относительная частота документов различных семантических единиц (df). Так, например, термин «объект» часто встречается в экономических статьях, в то время как слово «стратегия» - только в ограниченном количестве текстов. Причина заключается в том, что слово «объект» используется в целой группе различных контекстов, в то время как «стратегия» обозначает определённый термин. Часто используется общее преобразование, позволяющее отразить специфические особенности семантических единиц, а также общие их частоты. Такое преобразование называется *обратной частотой документа* (для *i*-го слова и *j*-го документа) [8]:

$$idf(i,j) = \begin{cases} 0, & \text{если } wf_{i,j} = 0; \\ \left(1 + log\left(wf_{i,j}\right)\right) log\,\frac{N}{df_i}, & \text{если } wf_{i,j} \geq 1. \end{cases}$$

В этой формуле N - общее число загруженных в систему текстов, а df_i - частота документов для i-й семантической единицы (число текстов, в которых встречается определённая группа слов). Данная формула содержит логарифмическую частоту слова, а также взвешивающий фактор, который равен 0, если слово появилось во всех документах, и максимален по значению, если семантическая единица появилась только в одном документе. Подобный подход позволяет формировать индексы, отражающие относительную частоту появления слов в тексте, а также точнее отразить их семантический смысл в обрабатываемом массиве данных. Таким образом, интеллектуальный анализ текстов позволяет не только находить закономерности в отдельных документах, но и строить гипотезы о зависимостях между различными

новостями, загруженными в автоматизированную систему [8].

Теперь рассмотрим заключительный элемент предлагаемой информационной системы, а именно модуль корреляционного анализа котировок фондового рынка. В общем случае это группа математических методов, позволяющих определить наличие зависимости между несколькими величинами. Как часть системы интеллектуального анализа данных, модуль корреляционного анализа позволяет в соответствии с загруженными в систему данными получить информацию о взаимосвязи отдельных объектов на основании их числовых характеристик.

В общем случае выделяют положительную корреляцию (при увеличении одного параметра второй также увеличивается) и отрицательную (обратная ситуация). Обнаруженной между объектами взаимосвязи необходимо найти численное выражение, в противном случае полученные данные будет сложно использовать в дальнейшем для нахождения закономерностей в массиве данных. Обычно для этого вводится так называемый *коэффициент корреляции*. Представим методику его расчёта.

Есть массив из n точек $\{x_{1,i}, x_{2,i}\}$, для которых рассчитываются средние арифметические [9]:

$$\overline{x_1} = \frac{\sum x_{1,i}}{n}, \overline{x_2} = \frac{\sum x_{2,i}}{n}.$$

Затем для полученных значений рассчитывается коэффициент корреляции:

$$r = \frac{\sum (x_{1,i} - \overline{x_1})(x_{1,i} - \overline{x_2})}{\sqrt{\sum (x_{1,i} - \overline{x_1})^2} \sqrt{\sum (x_{2,i} - \overline{x_2})^2}},$$

где r изменяется в пределах [-1;1]. В нашем случае представлен линейный коэффициент корреляции, который демонстрирует степень линейной взаимосвязи между параметрами x_1 и x_2. Коэффициент равен 1 или -1, если связь между величинами максимально близка к линейной.

Для коэффициента корреляции выдвигаются некоторые гипотезы, которые позволяют собирать информацию о входном массиве данных. Самыми распространёнными являются следующие:

1. Между величинами существует взаимосвязь (коэффициент значимо отличается от 0). Тестовая статистика в этом случае вычисляется по формуле [9]

$$\varepsilon = \left(0.5 \ln\left(\frac{1+r}{1-r}\right) - \frac{|r|}{2(n-1)}\right)\sqrt{n-3}$$

и сравнивается с табличным значением коэффициента Стьюдента: $t(p = 0{,}95, f = \infty) = 1{,}96$.

Если тестовая статистика больше табличного значения, то коэффициент значительно отличается от нуля. По формуле видно, что чем больше проведено измерений n, тем точнее полученный результат.

2. Отличие между двумя коэффициентами корреляции значимо. В данном случае вычисляется тестовая статистика [9] которая также сравнивается с табличным значением коэффициента Стьюдента $t(p,\infty)$.

$$\varepsilon = 0{,}5 \ln\left(\frac{(1+r_1)(1-r_2)}{(1-r_1)(1+r_2)}\right) \frac{1}{\sqrt{\frac{1}{n_1-3} - \frac{1}{n_2-3}}},$$

В результате применения подобных математических методов можно делать предположение о связи значений котировок на фондовом рынке. Так, высокий коэффициент корреляции позволяет говорить о том, что изменение котировок акций одного эмитента положительно (или отрицательно – в зависимости от характера корреляции) влияет на изменение ценовых характеристик другого. Таким образом, интеллектуальный анализ данных позволит более точно прогнозировать будущие состояния объектов и формировать гипотезы об их взаимных отношениях. На рис. 2 представлен алгоритм работы модуля корреляционного анализа временных рядов статистических данных фондового рынка.

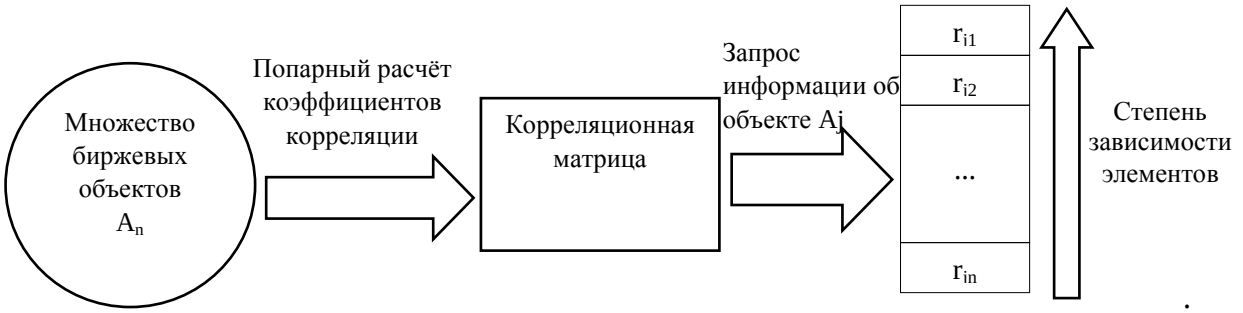

Рис. 2. Корреляционный анализ котировок фондового рынка

Комплексный подход к обработке статистической информации о фондовых рынках помогает решить проблему так называемых больших данных (Big Data) [10]. Представление информации в доступном любому пользователю виде упрощает получение информации в данной экономической сфере людям, далёким от фондовых рыков, а методы и модели компьютерного интеллекта позволяют проводить углублённый анализ информации без привлечения специалистов.

Функциональную схему программного комплекса «Фондовая аналитика» можно увидеть на рис. 3, где в виде блок-схемы обозначены основные этапы работы предлагаемой системы. Особенное внимание уделено распределению вычислений на несколько параллельных потоков при проведении интеллектуального анализа данных.

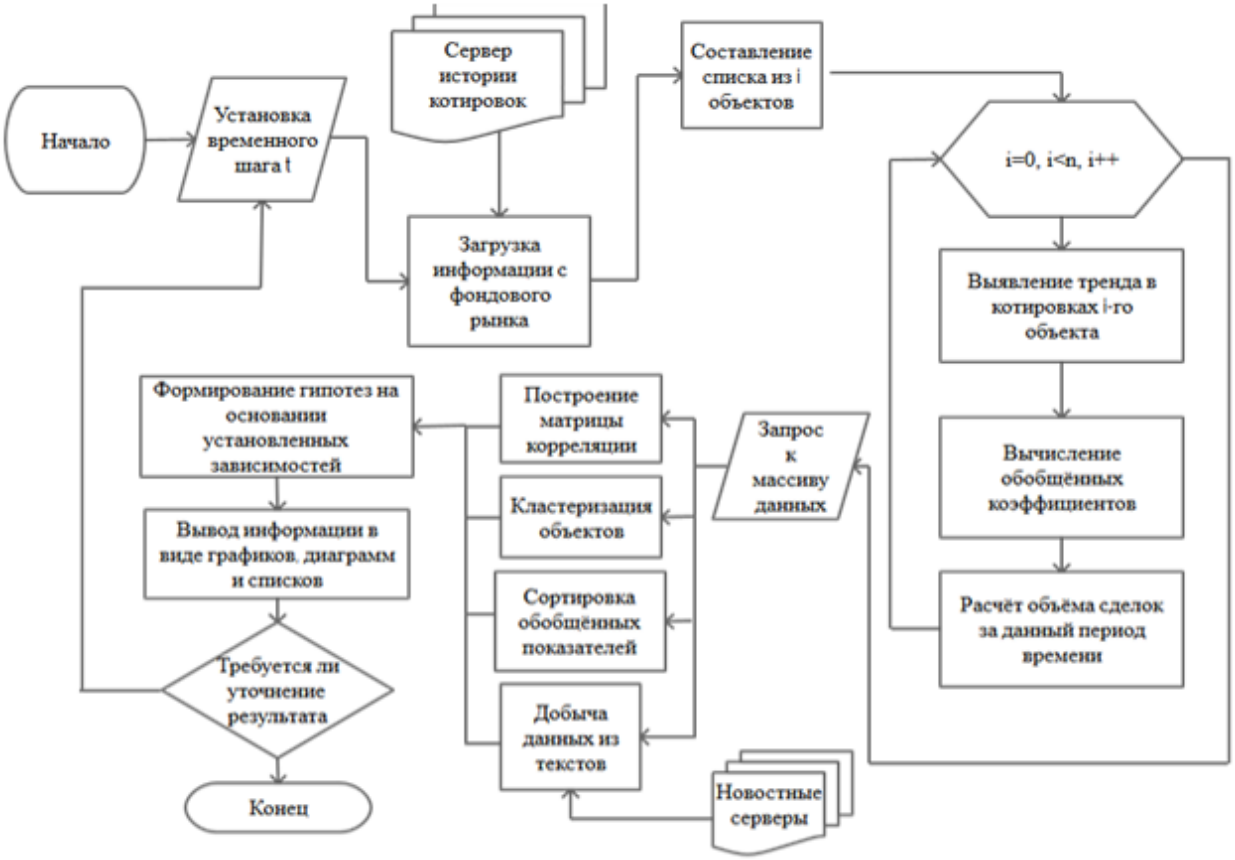

Рис. 3. Блок-схема работы автоматизированной системы «Фондовая аналитика»

Основным преимуществом данной автоматизированный системы является возможность аккумулирования значительных объёмов информации, их переработка и представление в виде сводных критериев и обобщённых зависимостей. Подобный подход может эффективно использоваться для решения управленческих задач в сфере

инвестирования, размещения государственных заказов и организации инновационных проектов. Предоставление результатов работы математических моделей в виде упорядоченного массива данных также при необходимости упростит интеграцию комплекса в другие информационные системы.

СПИСОК ЛИТЕРАТУРЫ

1. Барсегян, А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А.А. Барсегян, М.М. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.: БХВ-Петербург, 2007. – 384 с.
2. Чубукова, И.А. Data Mining / И.А. Чубукова. – М.: Бином. Лаборатория знаний, 2010. – 382 с.
3. Катаева, Е.С. Макростатистический анализ и прогнозирование / Е.С. Катаева. - Томск, 2016. - 56 с.
4. Швагер, Дж. Технический анализ. Полный курс / Дж. Швагер. - М.: Альпина Паблишер, 2016. – 804 с.
5. Швецов, Я.А. Основные проблемы информационной поддержки принятия решений в процессе управления трейдинговой инвестиционной деятельностью / Я.А. Швецов // Устойчивое развитие регионов: материалы междунар. науч.-практ. конф.: в 5 т. - Тамбов: Изд-во ТГТУ, 2016. - Т. 1. - С. 197-202.
6. Швецов, Я.А. Информационная поддержка принятия решений при размещении государственных заказов с учетом динамики изменения котировок ценных бумаг / Я.А. Швецов, В.И. Аверченков // Информационные системы и технологии. - Орёл, 2017. - Вып. 6.
7. Weiss, S.M. Text Mining: Predictive Methods for Analyzing Unstructured Information / S.M. Weiss, T. Zhang. - Berlin: Springer, 2007. – 237 p.
8. Хенрик, Б. Машинное обучение / Б. Хенрик, Д. Ричардс, М. Феверолф. - СПб.: Питер, 2017. - 336 с.
9. Суслов, В.И. Эконометрия / В.И. Суслов, Н.М. Ибрагимов, Л.П. Талышева, А.А. Цыплаков. - Новосибирск: СО РАН, 2005. - 744 с.
10. Майер-Шенбергер, В. Большие данные / В. Майер-Шенбергер, К. Кукьер; пер. с англ. И. Гайдюк. - М.: МИФ, 2014. - 240 с.

1. Barsegyan, A.A. *Techniques of Data Analysis*: Data Mining, Visual Mining, Text Mining, OLAP / A.A. Barsegyan, M.M. Kupriyanov, V.V. Stepanenko, I.I. Kholod. – S-Pb.: BHV-Petersburg, 2007. – pp. 384.
2. Chubukova, I.A. Data Mining / I.A. Chubukova. – M.: Binomial. Knowledge Laboratory, 2010. - pp. 382.
3. Kataeva, E.S. *Macro-statistical Analysis and Forecasting* / E.S. Kataeva. – Tomsk, 2016. – pp. 56.
4. Schwager, J. *Technical Analysis. Complete Course* / J. Schwager. – M.: Alpina Publisher, 2016. – pp. 804.
5. Shvetsov, Ya.A. Basic problems in information support of decision making during trading investment activity management / Ya.A. Shvetsov // *Stable Development of Regions: Proceedings of the Inter. Scientif. Pract. Conf.*: in 5 Vol. – Tambov: TSTU Publishers, 2016. – Vol.1. – pp. 197-202.
6. Shvetsov, Ya.A. Information support in decision making at state order placements taking into account dynamics in quotation of securities / Ya.A. Shvetsov, V.I. Averchenkov // *Information Systems and Technologies*. – Orel, 2017. – Issue 6.
7. Weiss, S.M. Text Mining: Predictive Methods for Analyzing Unstructured Information / S.M. Weiss, T. Zhang. - Berlin: Springer, 2007. – 237 p.
8. Henrik, B. *Machine Training* / B. Henrik, D. Richards, M. Feverolf. – S-Pb.: Peter, 2017. – pp. 336.
9. Suslov, V.I. *Econometrics* / V.I. Suslov, N.M. Ibragimov, L.P. Latysheva, A.A. Tsyplakov. - Novosibirsk: SB of RAS, 2005. – pp. 744.
10. Maier-Schoenberger, W. *Large Data* / W. Maier-Schoenberger, K. Coukier: transl. from Engl. I. Gaidyuk. – MIF, 2014. – pp. 240.

Статья поступила в редколлегию 20.11.17.
Рецензент: д.пс.н., профессор Брянского государственного технического университета
Спасенников В.В.

Сведения об авторах:

Швецов Ярослав Александрович, аспирант кафедры «Компьютерные технологии и системы» Брянского государственного технического университета, тел.: 8-953-289-97-64, e-mail: yshvetsov1491@yandex.ru.

Аверченков Владимир Иванович, д.т.н., профессор кафедры «Компьютерные технологии и системы» Брянского государственного технического университета, тел.:(4832) 56-05-33, e-mail: aver@tu-bryansk.ru.

Рытов Михаил Юрьевич, к.т.н., доцент, зав. кафедрой «Системы информационной безопасности» Брянского государственного технического университета, тел.: (4832) 51-13-77, e-mail: rmy@tu-bryansk.ru.
Фёдоров Владимир Павлович, д.т.н., профессор кафедры «Технология машиностроения» Брянского государственного технического университета, тел.: (4832) 58-82-20 e-mail: tm-bgtu@yandex.ru.
Федяева Галина Анатольевна, д.т.н., доцент кафедры «Электронные, радиоэлектронные и электротехнические системы» Брянского государственного технического университета, тел.: (4832) 56-36-02, e-mail: aep-bgtu@yandex.ru.

Shvetsov, Yaroslav Alexandrovich, Post graduate student of the Dep. “Computer Techniques and Systems”, Bryansk State Technical University, e-mail: yshvetsov1491@yandex.ru.
Averchenkov Vladimir Ivanovich, D. Eng., Prof. of the Dep. “Computer Techniques and Systems”, Bryansk State Technical University, e-mail: aver@tu-bryansk.ru.
Rytov Mikhail Yurievich, Can. Eng., Assistant Prof., Head of the Dep. “Information Safety Systems”, Bryansk State Technical University, e-mail: rmy@tu-bryansk.ru.
Fyodorov Vladimir Pavlovich, D. Eng., Prof. of the Dep. “Engineering Techniques”, Bryansk State Technical University, e-mail: tm-bgtu@yandex.ru.
Fedyaeva Galina Anatolievna, D. Eng., Assistant Prof. of the Dep. “Electronic, Radio-Electronic and Electro-Technical Systems”, Bryansk State Technical University, e-mail: aep-bgtu@yandex.ru.