DOI: 10.17323/2587-814X.2022.1.42.55

Информационно-логическая модель экспресс-анализа соответствия состояния предприятия, удовлетворяющего нормативам и регламентам, на основе общедоступных данных

Т.К. Богданова
E-mail: tanbog@hse.ru

Л.В. Жукова
E-mail: lvzhukova@hse.ru

Национальный исследовательский университет «Высшая школа экономики»
Адрес: Россия, 101000, г. Москва, ул. Мясницкая, д. 20

Аннотация

В последние 10 лет наблюдается взрывной рост объемов информации, размещаемых в сети интернет и цифровой экономики, формирование официальных баз данных различных государственных органов власти. Наличие большой информационной базы, открытой для исследования, способствует развитию новых методов и подходов к решению аналитических задач. Построение систем управления и поддержки принятия решений на основе использования объединенных в единое целое разрозненных открытых источников данных позволяет конечным пользователям принимать наиболее эффективные решения. Именно такой подход является основой для роста бизнеса компаний и повышения уровня зрелости управленческой деятельности на всех уровнях, альтернативы ему на данный момент нет. Такой подход, в конечном итоге, создает условия для дальнейшего роста экономики в целом. В работе предлагается информационно-логическая модель экспресс-анализа соответствия социально-экономического состояния предприятия нормативным требованиям со стороны контрольных и надзорных органов на основе открытой общедоступной информации. Выводы, сделанные на основе проведенного экспресс-анализа, служат обоснованием для принятия решения о необходимости более детального, углубленного анализа состояния отдельных предприятий.

Ключевые слова: экспресс-анализ, информационно-логическая модель, состояние предприятия, нормативные требования, контрольные и надзорные органы, открытые источники информации, общедоступные данные, структурированные и неструктурированные данные

Цитирование: Богданова Т.К., Жукова Л.В. Информационно-логическая модель экспресс-анализа соответствия состояния предприятия, удовлетворяющего нормативам и регламентам, на основе общедоступных данных // Бизнес-информатика. 2022. Т. 16. № 1. С. 42–55. DOI: 10.17323/2587-814X.2022.1.42.55

Введение

В работе предлагается информационно-логическая модель экспресс-анализа соответствия социально-экономического состояния предприятия нормативным требованиям со стороны контрольных и надзорных органов на основе общедоступной информации. Информационно-логическая модель построена на основе предложенной концепции, одной из важных особенностей которой является то, что концепция учитывает любые требования разных регуляторов, как количественных, так и качественных, предъявляемых к экономическим объектам разных типов (предприятиям, организациям, образовательным учреждениям и т. п.) [1]. Для разных типов предприятий и требований, предъявляемых к ним со стороны регулятора, необходимо на основе общедоступной информации формировать разные наборы компонент, в результате агрегирования которых с помощью разработанной таблицы поиска рассчитывается значение интегрального показателя, являющегося основой экспресс-анализа. Каждая компонента характеризует различные аспекты деятельности предприятия: экономические, социальные, финансовые, технические и т. п., и оценивается в соответствии с методами машинного обучения, математической статистики и эконометрики [2, 3].

Для проведения экспресс-анализа состояния предприятия используется как структурированная, так и неструктурированная информация. Неструктурированная информация предварительно структурируется с использованием различных методов обработки текстовой информации [4–6].

Динамика окружающей среды возрастает, стабильность внешней среды снижается, при этом требования к быстрому реагированию на кризисы растут. Количество информации, которую необходимо обработать для принятия того или иного решения, последовательно увеличивается, одновременно ужесточаются требования к качеству, безопасности и актуальности этой информации.

Одновременное использование структурированных и неструктурированных статистических данных позволяет получить более точную качественную оценку объекта исследования с учетом изменений, которые еще не отражены в официальной статистической отчетности, предоставляемой с определенной периодичностью и неизбежным запаздыванием.

Результатом проведенного экспресс-анализа является оценка соответствия социально-экономического состояния предприятия нормативным требованиям со стороны регулятора. Выводы, сделанные на основе проведенного экспресс-анализа, служат обоснованием для принятия решения о необходимости более детального, углубленного анализа отдельных предприятий.

Среди научных работ последних лет, посвященных различным экономическим и математическим исследованиям, все чаще делается акцент на применение современных цифровых технологий обработки больших объемов структурированных, слабоструктурированных и неструктурированных данных из открытых источников сети интернет, методов машинного обучения и искусственного интеллекта в моделях поддержки принятия решений [7–10].

Использование инновационных цифровых возможностей сбора и анализа общедоступной информации из сети интернет позволяет дополнительно анализировать характеристики качества работы различных предприятий и других объектов исследования. Подобный анализ на открытых данных можно проводить при помощи вспомогательного независимого инструментария оценки состояния объекта исследования, созданного на основе анализа больших объемов структурированных, слабоструктурированных и неструктурированных данных из открытых источников сети интернет, и сравнивать результаты с официальной методикой исследования по внутренним или официальным статистическим данным.

Контрольные меры, принятые на основе официальной статистической информации, могут поступать с большим запаздыванием, ведь между окончанием отчетного периода и передачей официальных статистических данных о состоянии объекта в органы государственной власти может проходить от 3 до 8 месяцев, что затрудняет оперативное реагирование в форс-мажорных ситуациях.

В научных исследованиях многими авторами предлагаются различные экономико-математические модели, базирующиеся на официальной статистической информации [11]. Большинство из них представляют собой эконометрические модели или модели, использующие методы машинного обучения. Как правило, имеющиеся статистические данные разбиваются на группы (демографические, социальные, финансовые и т.д.), ранжируются, или каким-то образом объединяются в единый интегральный показатель, а факторам присваиваются веса. Зачастую результатом подобного исследования становится интегральный показатель (коэффициент), который удобен при сопоставлении объектов.

Подобные инструменты в значительной степени опираются на внутренние данные или на имеющуюся статистическую базу [12].

Использование результатов анализа структурированных и неструктурированных данных из открытых источников сети интернет является наиболее полным и разносторонним способом для полноценного всестороннего анализа состояния экономического объекта исследования. Это позволяет получать объективную информацию о текущей ситуации, не подвергнутую промежуточной обработке и полученную на основании анализа большого количества разнообразных актуальных данных, которые хранятся в открытом доступе в любых источниках сети интернет. При необходимости результаты анализа внешних данных могут быть соотнесены с результатами аналогичной аналитической деятельности, проведенной на основе использования внутренних данных. Также результаты анализа по открытым общедоступным источникам могут дополнять официальные или внутренние данные в некоторых аспектах деятельности объекта исследования.

Преимущество использования открытых данных – это возможность получать информацию с любой периодичностью (не привязываясь к регулярности обновления, официально публикуемой статистической отчетности), расширять и проверять соответствие фактического социально-экономического состояния объекта исследования официальным данным.

1. Классификация источников общедоступной информации

Вся общедоступная информация может быть представлена в виде данных различного типа. В настоящее время все существующие данные можно разделить на:

1) структурированные;
2) слабоструктурированные;
3) квазиструктурированные;
4) неструктурированные.

К структурированным данным относятся данные, которые упорядочены определенным способом, имеют заданную структуру, и описывают конкретную предметную область. В совокупности это позволяет проводить достоверный и глубокий анализ этих данных. Чаще всего такая информация представлена в виде таблиц.

К слабоструктурированным данным можно отнести те данные, которые не соответствуют четкой структуре таблиц и отношений в БД, но при этом содержат в себе специальные разграничители (теги), которые позволяют семантически разделять весь объем данных. В качестве примера можно привести XML-документы.

К квазиструктурированным данным относятся данные с неустойчивым форматом, требующие для своей обработки специальными инструментами больших временных затрат. Примером таких данных может служить страница сайта.

К неструктурированным данным можно отнести данные, которые не имеют определенной формы и не являются строго зафиксированными. В настоящий момент такой формат данных является преобладающим в связи с развитием информатизации населения, примерно 80% всей имеющейся на данный момент информации, являются неструктурированной. Примером таких данных являются изображения, видео, аудио и текстовая информация из социальных сетей.

Данные в зависимости от их типа требуют своих методов предобработки и обработки. Большинство методов математической статистики и эконометрики основаны на анализе структурированной информации. Методы машинного обучения, нейронные сети позволяют анализировать слабоструктурированные, квазиструктурированные и неструктурированные данные, выделяя в них закономерности. Также при проведении различных процедур предобработки эти данные могут быть сведены к структурированным и включены в классические математические модели.

Если неструктурированные данные представлены текстом, то предварительная их обработка с помощью методов векторизации и классификации позволяет привести их к структурированному виду.

Информация для формирования базы исследования для проведения экспресс-анализа может быть получена из различных источников, отличающихся статусом, частотой обновления и степенью достоверности предоставляемой информации. В *таблице 1* представлена классификация источников общедоступной информации с учетом степени надежности источника.

Предложенная в данной статье информационно-логическая модель построения интегрального показателя для экспресс-анализа соответствия социально-экономического состояния предприятия нормативным требованиям со стороны контрольных и надзорных органов, базируется на концептуальной модели экспресс-анализа изложенной в [1]. Отличи-

Таблица 1.

Классификация источников общедоступной информации

Источник информации	Характеристика источника информации	Пример источника информации	Тип информации	Обновление данных на источнике
Официальные генераторы и агрегаторы данных	Сайты федерального и региональных органов статистики, сайты министерств и ведомств, публикующие по положению о раскрытии информации тематические данные, достоверность которых подтверждается соответствующим органом государственной власти.	rosstat.gov.ru zakupki.gov.ru fssp.gov.ru cbr.ru wciom.ru	Структурированные данные.	Как правило, периодичность обновления 1 раз в квартал, или реже.
Сайты и страницы в соц. сетях объектов исследования	Сайты предприятий, организаций всех форм собственности, сайты площадок, на которых они обязаны размещать информацию о своей деятельности. Достоверность информации, как правило, подтверждается только самим объектом исследования.	technomoscow.ru uniconf.ru tinkoff.ru 57.mskobr.ru	Все типы данных.	Постоянное обновление.
Неофициальные генераторы данных	Сайты организаций, занимающихся деятельностью, связанной с объектами исследования и публикующие данные о них в открытых источниках. Достоверность обеспечивается внутренним мониторингом и контролем информации.	cian.ru hse.ru/rlms	Преимущественно структурированные данные.	Согласно утвержденной методике обновление может проводиться как с заданной периодичностью, так и на постоянной основе.
Неофициальные агрегаторы данных	Российские и международные агрегаторы данных, обычно предоставляющие их для проведения научных и иных исследований. Достоверность обеспечивается внутренним мониторингом.	bankodrom.ru banki.ru avtostat.ru data.worldbank.org	Преимущественно структурированные данные.	Обычно обновление проводится с периодичностью, соответствующей периодичности обновления официальных данных.
Неофициальные интернет–источники экспертных исследований	Российские и международные сайты экспертных организаций, рейтинговых агентств, персональные страницы признанных экспертов. Достоверность данных обеспечивается репутацией эксперта.	raexpert.ru ra–national.ru	Все типы данных.	Обновление проводится согласно внутренним правилам источника информации.
Неофициальные общедоступные интернет–источники	Страницы в социальных сетях, блоги, комментарии под контентом, страницы неофициальных сообществ.Достоверность данных, как правило, не подлежит проверке.	moneyzz.ru pedsovet.su	Преимущественно неструктурированные или абоструктурированные данные.	Постоянное обновление.

тельной особенностью предложенной концептуальной модели является то, что в качестве отправной точки авторами предлагается учитывать требования контролирующих органов, в то время как в большинстве российских и зарубежных исследований оценка состояния объекта исследования осуществляется исходя из требований, предъявляемых к объекту его владельцами или инвесторами. Еще одно преимущество концептуальной модели – это использование общедоступных данных, то есть возможность получать информацию в любой момент времени, не привязываясь к периодам обновления официально публикуемой статистической отчетности, и возможность проверять соответствие фактического состояния объекта исследования официальным данным. Предлагаемая информационно-логическая модель представляет собой объединение алгоритма расчета отдельных компонент интегрального показателя с применением математических, эконометрических и статистических методов, характеристик входной и выходной информации на каждом этапе, и, собственно, алгоритма расчета значений интегрального показателя с помощью логической функции на основе таблицы поиска.

2. Компоненты интегрального показателя и методы их оценки

Интегральный показатель представляет собой гибкий инструментарий экспресс-анализа на основе общедоступных структурированных и неструктурированных данных. Алгоритм построения интегрального показателя основан на агрегировании отдельных значений каждой компоненты из набора с применением таблицы поиска. Каждая компонента оценивается с использованием математических, эконометрических и статистических методов, таких как: логистическая регрессионная модель, методы кластеризации и группировки, методы тематического моделирования и т.д.

Гибкий инструментарий экспресс-анализа для принятия управленческих решений, разработанный на основе концептуальной модели, представляет собой последовательность из пяти этапов, начиная от требований со стороны контрольных и надзорных органов, разработки и оценки набора компонент, характеризующих объект исследования, их агрегирование в единый интегральный показатель на основе таблицы поиска, и заканчивая мониторингом и ранжированием объектов исследования по результатам расчетов [1].

В зависимости от типа объекта исследования (промышленное предприятие, банковская организация, образовательное учреждение и т.п.), исходя из требований различных регуляторов, формируется перечень источников данных для проведения экспресс-анализа: сайты объектов исследования, новостные источники, электронные площадки или агрегаторы информации, сайты государственных органов власти и т.д. На основе информации из этих источников создается база данных исследования. Гибкость предлагаемого в статье инструментария обусловлена тем, что перечень необходимых для проведения экспресс-анализа компонент может быть дополнен в зависимости от типа объекта исследования, от предъявляемых к объекту исследования часто меняющихся требований со стороны контрольных и надзорных органов и роста числа источников общедоступной информации.

В *таблице 2* приведен перечень выделяемых авторами возможных компонент, относящихся к четырем блокам типов входной информации для расчета компонент, типов переменных рассчитанного значения каждой компоненты (в соответствии с метриками, предложенными Робертом С. Капланом и Дэйвидом П. Нортоном), и методов оценки значений компонент [13].

Для оценивания компонент интегрального показателя на основе информации об объектах исследования из базы данных применяются различные методы оценивания.

2.1. Компонента 1. Вероятность финансового неблагополучия

Представляет собой вероятность наступления неблагополучного финансового состояния объекта исследования (банкротство, отзыв лицензии по финансовым причинам). Для оценивания этой вероятности применяется регрессионная логистическая модель на основе данных финансовой отчетности, а также показателей их волатильности: стандартного отклонения и дисперсии, данных макроэкономических переменных, данных о государственных закупках в качестве поставщика или покупателя. В общем виде логистическая регрессионная модель принимает вид [1]:

$$P\left(Y=1 \mid x,\, m,\, v\right)=\frac{1}{1+e^{-z}},$$

$$z=\beta_0+\sum\beta_i x_i+\sum\gamma_j m_j+\sum\varphi_k v_k, \text{ где:}$$

$P\left(Y=1 \mid x,\, m,\, v\right)$ – условная вероятность неблагополучного финансового состояния объекта исследования;

β_0 – константа;

x_i – переменные, характеризующие финансовое состояние объекта исследования;

m_j – переменные, характеризующие внешнюю по отношению к объекту исследования среду (макроэкономические факторы);

v_k – неколичественные показатели деятельности объекта исследования;

$\beta_i, \gamma_j, \varphi_k$ – коэффициенты регрессии, которые должны быть оценены.

2.2. Компоненты 2 и 3. Статус объекта исследования по масштабу и принадлежности к аномальной группе

Представляет собой результаты кластеризации по определению принадлежности объекта исследования к одному из классов. Эти компоненты позволяют учесть особенности всех объектов исследования данного типа по признакам местоположения, масштаба, вида деятельности и т.п. При этом учитывается специфика полученного кластера объектов,

Таблица 2.

Компоненты интегрального показателя и методы их оценки

№	Компоненты	Тип исходной информации	Тип переменной рассчитанного значения компоненты	Метод оценки
Характеристика финансового состояния объекта исследования				
1	Вероятность финансового неблагополучия	структурированная	категориальная, порядковая	логистическая регрессионная модель
Статусная идентичность объекта исследования				
2	Статус объекта исследования по масштабу	структурированная	категориальная	кластерный анализ
3	Статус объекта исследования по принадлежности к аномальной группе	структурированная	категориальная	кластерный анализ
Характеристика внешней информационной среды				
4	Медийная активность относительно объекта исследования	слабоструктурированные, квазиструктурированные и неструктурированные данные	количественная	семантический анализ
5	Положительная тональность упоминаний об объекте исследования в интернет–источниках		количественная	семантический анализ
6	Негативная тональность упоминаний об объекте исследования в интернет–источниках		количественная	семантический анализ
Нормативные требования к состоянию объекта исследования				
7	Соответствие требованиям государственных органов	структурированная	бинарная или категориальная	статистический и индексный анализ

что позволяет более объективно оценить состояние предприятия относительно объектов из его класса.

Алгоритмы кластеризации подразделяются на два типа:

1. Иерархические методы.

2. Неиерархические методы.

Методы иерархической кластеризации бывают двух видов [14, 15]:

1. Агломеративные (объединяющие).

В этой категории методов происходит объединение исходных объектов и уменьшение количества кластеров [16]. Такой подход осуществляется «снизу-вверх»: создание небольших кластеров и объединение их в более крупные.

2. Дивизивные (разъединяющие).

Для алгоритмов дивизивного вида характерно начальное условие наличия одного кластера. Этот изначальный кластер делится на более мелкие кластеры. Разъединяющие алгоритмы работают «сверху-вниз».

Недостатком этих методов является вычислительная сложность на данных большой размерности. Для иерархических методов кластеризации характерной особенностью является то, что наблюдения, попавшие однажды в кластер, при дальнейшем объединении (разъединении) объектов не могут переместиться в другой кластер, в отличие от неиерархических методов.

Основная отличительная идея неиерархических методов кластеризации – определение центра кластера и группировка всех объектов, находящихся на расстоянии от центра кластера в пределах заданного порогового значения [14, 15]. К группе методов неиерархической кластеризации относятся алгоритмы семейства *k*-means (*k*-среднее) [16].

Для данных большой размерности с неизвестным числом кластеров предлагается метод BIRCH (двухшаговая или двухступенчатая кластеризация), основанный на методе *k*-means. Двухступенчатая класте-

ризация не требует задания количества кластеров, так как на первом шаге определяется оптимальное количество кластеров, а затем уже происходит разбиение на однородные группы. Данный метод позволяет анализировать большие объемы как количественных, так и качественных данных, хорошо работает при небольших объемах памяти.

Качество полученной кластеризации может быть оценено с помощью силуэтной меры *Sil* [17]:

$$Sil = \frac{1}{N}\sum_{c_k \in C}\sum_{x_i \in c_k}\frac{b(x_i,c_k)-a(x_i,c_k)}{\max\left(a(x_i,c_k),b(x_i,c_k)\right)},$$

где:

Sil – общее значение силуэтной меры кластеризации всех данных;

N – общее количество объектов в выборке;

C – множество всех кластеров;

c_k – k-ый кластер на множестве C;

x_i – i-ый объект, $i \in [1, N]$;

$a(x_i,c_k) = \frac{1}{|c_k|}\sum_{x_i \in c_k}\left\|x_i - x_j\right\|$ –

среднее расстояние от объекта $x_i \in c_k$ до других объектов x_j из этого кластера c_k (компактность);

$|c_k|$ – количество объектов в кластере c_k;

$b(x_i,c_k) = \min_{c_l \in C \setminus c_k}\left\{\frac{1}{|c_l|}\sum_{x_j \in c_l}\left\|x_i - x_j\right\|\right\}$ –

среднее расстояние от объекта $x_i \in c_k$ до объектов x_j из другого кластера c_l: $k \neq l$, $k, l \in [1, C]$.

Силуэтная мера *Sil* принимает значения на отрезке от –1 до +1, при этом:

1 – все наблюдения расположены точно в центрах их кластеров;

-1 – все наблюдения расположены в центрах некоторых других кластеров;

0 – наблюдения расположены в среднем на равных расстояниях от центра их кластера и центра ближайшего кластера.

2.3. Компоненты 4, 5 и 6.

Медийная активность относительно объекта исследования, положительная и негативная тональность упоминаний об объекте исследования в интернет-источниках

Оценка этих компонент представляет собой анализ неструктурированных или слабоструктурированных данных, преимущественно текстовых. Для проведения семантического анализа по оценке значений компонент, характеризующих медийную активность и тональность упоминаний относительно объекта исследования, требуется предварительная предобработка этих данных, техническая и лингвистическая очистка данных, составление словаря употребляемых слов в текстах.

Тональность – это эмоциональное отношение автора к некоторому объекту, выраженное в тексте [18, 19]. Один из способов определения тональности – поиск эмоциональной составляющей в тексте по ранее сформированным тональным словарям с применением анализа лингвистики. Применение готовых словарей к очищенным текстовым данным позволяет классифицировать текстовые единицы (предложения, слова) на три категории: амбивалентное, положительное и отрицательное. Для семантического анализа медийной активности, категоризации текста и применения методов машинного обучения требуется векторизация текста.

Векторизация – это процесс конвертации текстовых документов в числовой вектор. Выбор метода векторизации, как правило, зависит от конкретного случая, условий, имеющихся аппаратных и технологических средств. Постоянное появление новых методов и алгоритмов, улучшающих качество векторизации и скорость обработки, позволяет внедрять процесс обработки естественного языка в модель.

В настоящий момент наиболее популярен реализованный во многих статистических пакетах алгоритм Bag-of-Words («мешок слов»). «Мешок слов» представляет собой векторное представление неупорядоченного набора слов в вектор размерности n [20–23]. Схематично алгоритм может быть представлен следующим образом.

Весь текст можно представить как набор обработанных слов, то есть отдельных термов ($_{tj}$), которые с помощью данного алгоритма переводятся в числовые данные из пространства R^n.

$$B: \text{words} \to R^n,$$

$$B(\textit{'some text in the Internet'}) = (w_{i,1}, w_{i,2}, ..., w_{i,n}),$$

где:

t_j – терм j;

w_{ij} – вес терма j в документе; вес документов нормируют так, чтобы $0 < w_{ij} < 1$, для $\forall i$;

n – количество термов в пространстве.

Документ в таком случае задаётся следующим образом:

$$d = (w_1, w_2, ..., w_{|V|}),$$

где:

d – вектор документа;

$|V|$ – количество уникальных термов в документе.

Вес терма можно задать несколькими вариантами:

1. Бинарным образом:

$$w_i = \begin{cases} 1, t_i \in d \\ 0, t_i \notin d \end{cases}.$$

2. По количеству вхождений терма:

$$w_i = n_i,$$

где n_i – количество вхождений терма в документ.

3. Частота терма (Term Frequency, TF):

$$w_i = tf(t_i, d) = \frac{n_i}{\sum_{k=1}^{|V|} n_k},$$

где:

tf – частота терма;

n_i – количество вхождений терма в документ;

$\sum_{k=1}^{|V|} n_k$ – количество термов в документе.

4. Частота терма – обратная частота документа (Term Frequency – Inverse Document Frequency, TF–IDF).

Представление в виде двух параметров: $w_{ij} = tf_i \cdot idf_i$, где tf_{ij} – это отношение числа термов t_i в документе d_j к общему числу термов в этом документе, а idf_i – число, обратное количеству документов, в котором встречается терм t_i. Таким образом, чем чаще слово встречается в этом документе, но реже встречается вообще во всех документах, тем больше вес этого терма в данном документе:

$$tf(t_i, d) = \frac{n_i}{\sum_{k=1}^{|V|} n_k},$$

$$idf(t, d) = \log \frac{|D|}{|d_i \supset t_i|},$$

где:

$d_i \supset t_i$ – количество документов, в которых встречается t_i;

$|D|$ – количество документов в корпусе.

Тогда вес считается следующим образом:

$$w_i = tf - idf(t_i, d, D) = tf(t_i, d) \cdot idf(t, d).$$

После векторизации применяются алгоритмы семантического анализа текста для определения тональности, главных тем, медийной активности и т.д.

Для расчета значений компонент применяются методы статистики для обобщения информации об объекте исследования, например, с помощью прямого подсчета встречаемости положительных и отрицательных слов определяется общая тональность текста.

2.4. Компонента 7.

Соответствие требованиям государственных органов

Эта компонента определяется как бинарный или порядковый показатель, рассчитанный с помощью индексов и статистических показателей. Он представляет собой оценку количества нарушений в деятельности объекта исследования, в случае если заданы нормативные и пороговые значения со стороны контрольных или надзорных государственных органов власти.

Консолидированным представлением вышеизложенного является информационно-логическая модель экспресс-анализа соответствия социально-экономического состояния объекта исследования требованиям контрольных и надзорных органов (*рис. 1*). На *рис. 1* этап 3, являющийся ключевым в алгоритме расчета интегрального показателя, показан в общем виде. Детализация этапа 3 информационно-логической модели представлена на *рис. 2*. В рамках этого этапа производится оценивание компонент интегрального показателя и расчет значений самого интегрального показателя в зависимости от значений каждой компоненты из набора.

Для преобразования значений компонент, характеризующих медийную активность (компонента 4), тональность упоминания об объекте исследования в интернет-источниках (компоненты 5 и 6) и соответствие требованиям государственных органов (компонента 7) предлагается использовать межквартильный размах IQR по выборке размера n. Здесь:

$F_n(x)$ – выборочная функция распределения;

$IQR = Q_3 - Q_1$, где $Q_3 = 0{,}75$; $Q_1 = 0{,}25$.

Предложенная информационно-логическая модель была апробирована на базе данных группы промышленных предприятий и предприятий финансовой сферы.

Был проведен экспресс-анализ соответствия потребности в финансовой помощи для 506 промышленных предприятий, зарегистрированных в г. Москве, и целесообразности ее оказания для федеральных и региональных органов власти. Экспресс-анализ проводился на основе открытых данных за 2016, 2017 и 2018 гг. Полученные результаты соот-

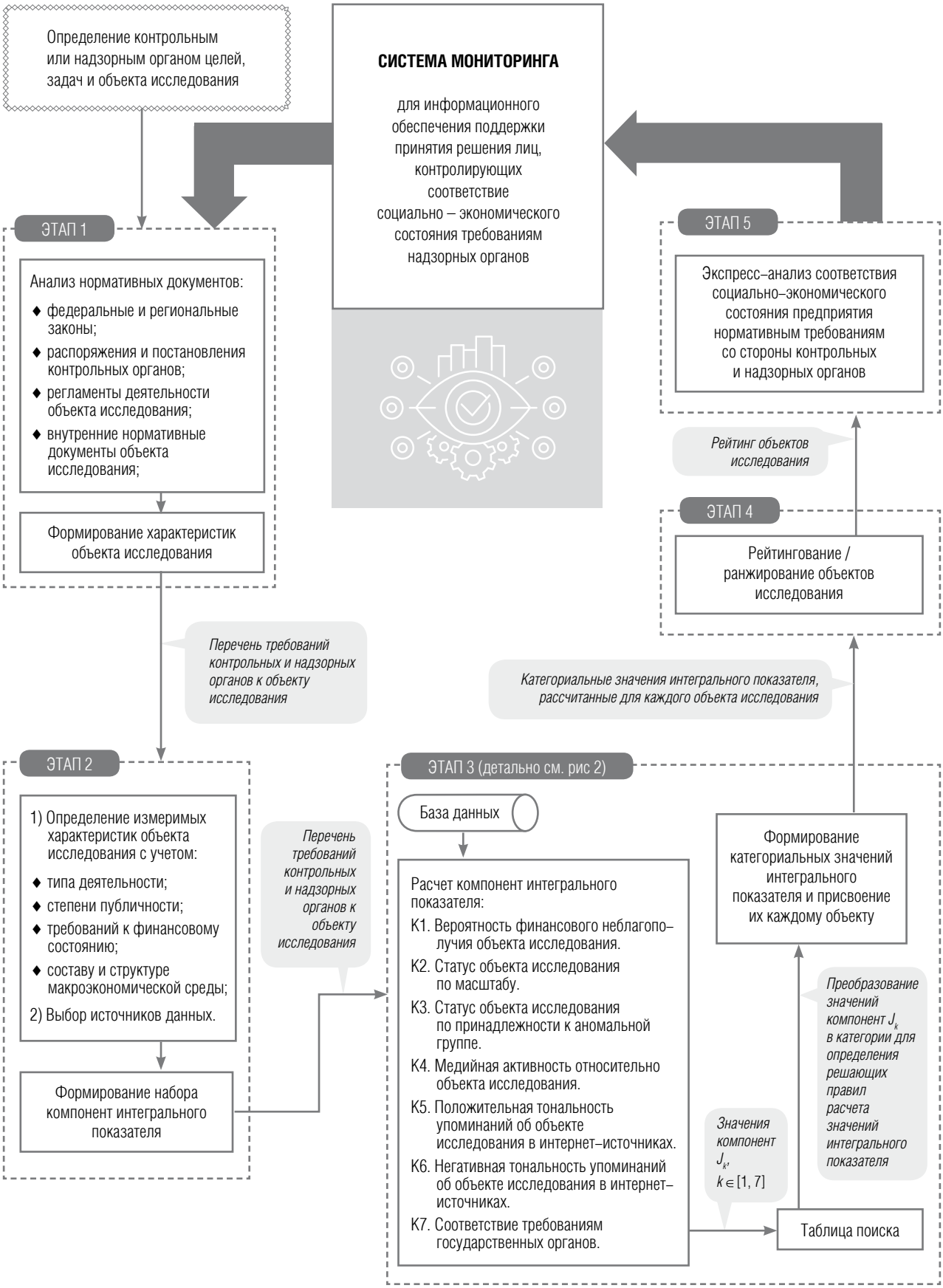

Рис. 1. Информационно–логическая модель алгоритма расчета компонент интегрального показателя.

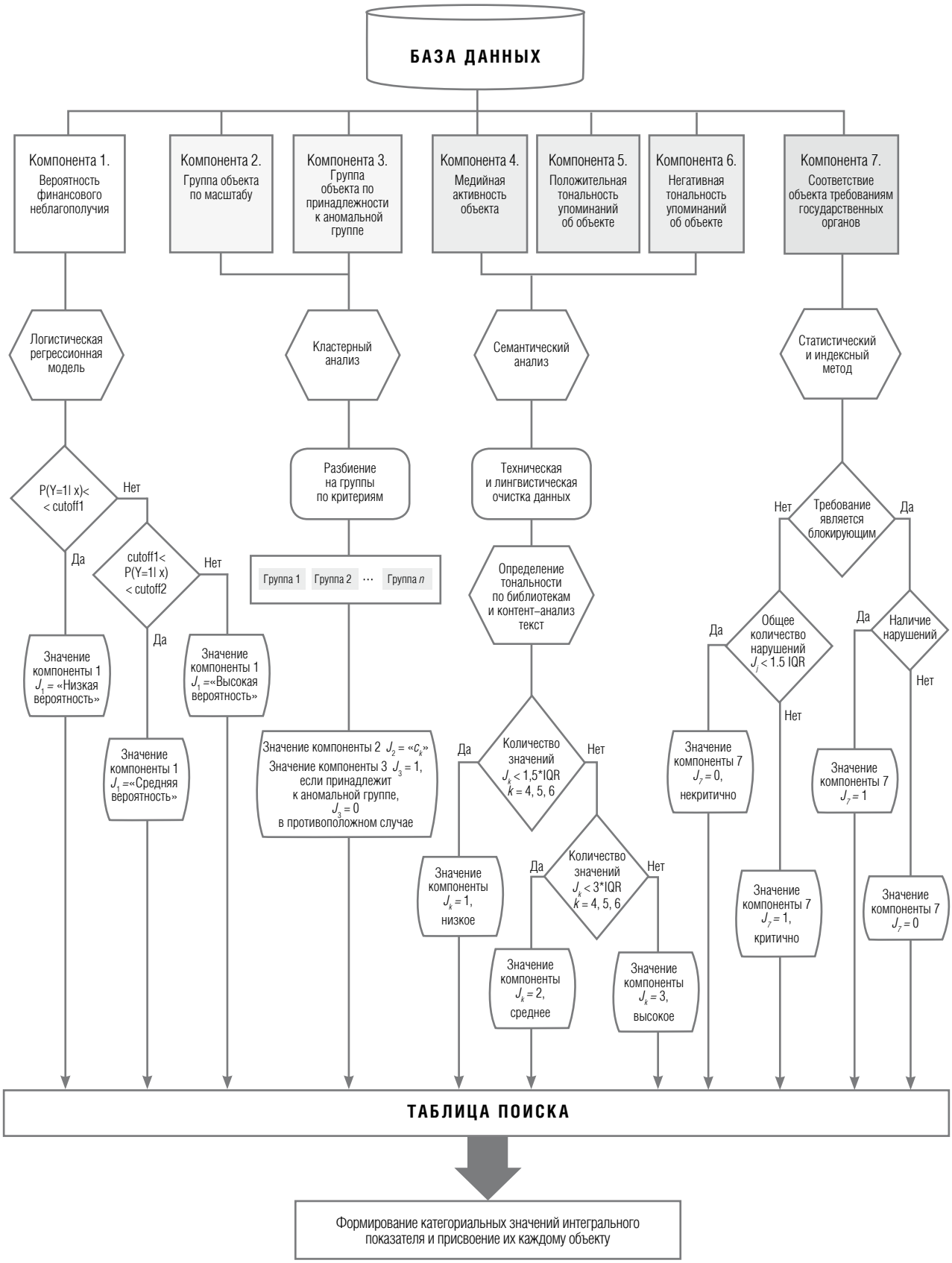

Рис 2. Детализация этапа 3 информационно–логической модели алгоритма расчета компонент интегрального показателя.

ветствовали фактическим данным на следующий год о назначении субсидий и льгот со стороны правительства г. Москвы [24].

Предложенная концептуальная модель проведения экспресс-анализа соответствия социально-экономического состояния объекта исследования заявленным требованиям со стороны контрольных и надзорных органов была апробирована для оценки социально-экономического состояния коммерческого банка. Контрольным органом в данном случае является ЦБ РФ – надзорный орган в банковской сфере. В соответствии с требованиями ЦБ РФ о надежности банка были получены значения четырех компонент интегрального показателя и рассчитано его значение для каждого банка. Прогностическая способность построенной модели была подтверждена их фактическим состоянием на март 2020 года [1].

Заключение

В статье предложена информационно-логическая модель экспресс-анализа соответствия социально-экономического состояния объекта требованиям со стороны контрольных и надзорных органов, с использованием открытых общедоступных данных. Предложенная информационно-логическая модель базируется на концепции использования интегрального показателя для экспресс-анализа соответствия социально-экономического состояния объекта независимо от его типа требованиям, предъявляемым к нему контрольными и надзорными органами.

Дана классификация информационных источников и методов их обработки в зависимости от типа данных.

Предложен алгоритм расчета выделяемых авторами возможных компонент, относящихся к четырем блокам типов входной информации, типов переменных рассчитанного значения каждой компоненты (в соответствии с метриками, предложенными Робертом С. Капланом и Дэйвидом П. Нортоном), и методов оценки значений компонент.

Разработанная концептуальная модель была апробирована для проведения экспресс-анализа соответствия социально-экономического состояния двух разных типов объектов предъявленным к ним требованиям со стороны надзорных органов на выборках из 506 промышленных предприятий [24] и 111 банков [1]. ■

Литература

1. Богданова Т.К., Жукова Л.В. Оценка состояния объекта управления на основе универсального комплексного индикатора с использованием структурированных и неструктурированных данных // Бизнес-информатика. 2021. Т. 15. № 2. С. 21–33. https://doi.org/10.17323/2587-814X.2021.2.21.33
2. Кричевский М.Л. Методы машинного обучения при выборе стратегии предприятия // Вопросы инновационной экономики. 2019. Т. 9. № 1. С. 251–266. https://doi.org/10.18334/vinec.9.1.40093
3. Опекунов А.Н., Кузьмина М.Г. Принципы формирования моделей прогнозирования вероятности банкротства предприятий с использованием элементов машинного обучения // Модели, системы, сети в экономике, технике, природе и обществе. 2019. № 4. С. 24–31.
4. Касевич В.Б. Элементы общей лингвистики. М.: Наука, 1977.
5. Савенков П.А. Использование методов и алгоритмов машинного обучения в системах поддержки принятия управленческих решений обучения // Вестник науки и образования. 2019. №1–2 (55). С. 23–25. https://doi.org/10.24411/2071-6168-2019-10207
6. Попова С.В., Ходырев И.А. Извлечение ключевых словосочетаний обучения // Научно-технический вестник информационных технологий, механики и оптики. 2012. №1 (77). С. 67–71.
7. Елисеева Е.Н. Финансовый инструментарий оценки несостоятельности промышленных предприятий обучения // Регион: системы, экономика, управление. 2019. № 3 (46). С. 132–140.
8. Медведев Д.А. Большие данные: причины появления и как их можно использовать обучения // Наука и образование сегодня. 2019. № 4 (39). С. 14–16.
9. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. 2008. Vol. 2. Nos. 1–2. P. 1–135. https://doi.org/10.1561/1500000011
10. Морозов А.Н. Альтернативные источники статистической информации как основа принятия политических решений обучения // Вопросы государственного и муниципального управления. 2018. № 2. С. 50–70.
11. Пузанов А.С., Трутнев Э.К., Маркварт Э., Попов Р.А., Сафарова М.Д. Стратегическое планирование и градорегулирование на муниципальном уровне. М.: Издательский дом «Дело» РАНХиГС, 2017.

12. Андреева Н.А., Угримова С.Н. К вопросу применения статистических методов интегральной оценки эффективности системы управления промышленными предприятиями обучения // Учет и статистика. 2019. № 1 (53). С. 42–49.

13. Нортон Д.П., Каплан Р.С. Сбалансированная система показателей. М: Олимп-Бизнес, 2008.

14. Чугунов В.Р., Жукова Л.В., Ковальчук И.М., Ковалева А.С. Математические методы группирования данных для принятия управленческих решений в задачах планирования обучения // Actual Problems of System and Software Engineering 2017. Proceedings of the 5th International Conference on Actual Problems of System and Software Engineering Supported by Russian Foundation for Basic Research. 2017. Project #17-07-20565. С. 333–341.

15. Баресягин А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А.А. Баресягин, М.С. Куприянов, В.В. Степаненко, И.И. Холод. 2-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2007.

16. Шалымов Д.С. Алгоритмы устойчивой кластеризации на основе индексных функций и функций устойчивости обучения // Стохастическая оптимизация в информатике. 2018. Т. 4. С. 236–248.

17. Kaufman L., Rousseeuw P. Finding groups in data: An introduction to cluster analysis. Wiley-Interscience, 2005.

18. Семина Т.А. Анализ тональности текста: современные подходы и существующие проблемы // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6, Языкознание: Реферативный журнал. 2020. № 4. С. 47–64.

19. Liu B. Sentiment analysis and subjectivity. In Handbook of Natural Language Processing, Second Edition (eds. N. Indurkhya, F.J. Damerau). London: Chapman and Hall/CRC, 2010.

20. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Е.И. Большакова и [др.]. М.: Изд-во НИУ ВШЭ, 2017.

21. Демидова Л.А., Степанов М.А. Подход к решению задачи выявления структурных трансформаций в группах временных рядов обучения // Cloud of science. 2019. № 2. С. 201–226.

22. Попова С.В., Ходырев И.А. Извлечение ключевых словосочетаний обучения // Научно-технический вестник информационных технологий, механики и оптики. 2012. №1 (77). С. 67–71.

23. Краснянский М.Н. Обухов А.Д., Соломатина Е.М., Воякина А.А. Сравнительный анализ методов машинного обучения для решения задачи классификации документов научно-образовательного учреждения обучения // Вестник Воронежского государственного университета. 2018. № 3. С. 173–182. https://doi.org/10.17308/sait.2018.3/1245

24. Жукова Л.В. Экспресс-анализ состояния промышленных предприятий Москвы с использованием универсального комплексного индикатора. М: Экономическая наука современной России. 2021. Т 4 (95). С. 89–96. https://doi.org/10.33293/1609-1442-2021-4(95)-89-97

Об авторах

Богданова Татьяна Кирилловна

кандидат экономических наук

доцент департамента бизнес-информатики, Высшая школа бизнеса, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: tanbog@hse.ru

ORCID: 0000-0002-0018-2946

Жукова Людмила Вячеславовна

старший преподаватель департамента прикладной экономики, Факультет экономических наук, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: lvzhukova@hse.ru

ORCID: 0000-0003-1647-5337

Information-logical model of express analysis of the state of the enterprise that meets the requirements of standards and regulations, based on publicly available data

Tatiana K. Bogdanova
E-mail: tanbog@hse.ru

Liudmila V. Zhukova
E-mail: lvzhukova@hse.ru

HSE University
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract

The last 10 years have witnessed an explosive growth in the volume of information posted on the Internet and the digital economy, as well as the formation of official databases of various public authorities. The availability of a large information base open for research has facilitated the development of new methods and approaches to solving analytical problems. Building management and decision-making support systems based on the use of united disparate open data sources allows end users to make the most effective decisions. This is the approach that underpins business growth and managerial maturity at all levels – there is no alternative. Such an approach ultimately creates the conditions for further growth of the economy as a whole. This paper proposes the information and logical model of express analysis of compliance of socio-economic condition of the enterprise with the regulatory requirements of the control and supervisory authorities on the basis of open, publicly available information. The conclusions drawn on the basis of express analysis serve as a basis for deciding on the need for a more detailed, in-depth analysis of the state of individual enterprises.

Keywords: express analysis, information-logical model, enterprise status, regulatory requirements, control and supervisory bodies, open sources of information, publicly available data, structured and unstructured data

Citation: Bogdanova T.K., Zhukova L.V. (2022) Information-logical model of express analysis of the state of the enterprise that meets the requirements of standards and regulations, based on publicly available data. *Business Informatics*, vol. 16, no. 1, pp. 42–55. DOI: 10.17323/2587-814X.2022.1.42.55

References

1. Bogdanova T.K., Zhukova L.V. (2021) The concept for valuation the position of the control object based on a universal complex indicator using structured and unstructured data. *Business Informatics*, vol. 15, no. 2, pp. 21–33 (in Russian). http://doi.org/10.17323/2587-814X.2021.2.21.33
2. Krichevskiy M.L. (2019) Methods of machine learning in choosing a strategy of an enterprise. *Russian Journal of Innovation Economics*, vol. 9, no. 1, pp. 251–266 (in Russian). https://doi.org/10.18334/vinec.9.1.40093
3. Opekunov A.N., Kuzmina M.G. (2019) Principles of forming models for forecasting the probability of bankruptcy of enterprises using machining elements. *Models, Systems, Networks in Economics, Technology, Nature and Society*, no. 4, pp. 24–31 (in Russian).
4. Kasevich V.B. (1977) *Elements of general linguistics*. Moscow: Nauka (in Russian).
5. Savenkov P.A. (2019) Using methods and algorithms of machine learning in management decision support systems. *Bulletin of Science and Education*, nos. 1–2 (55), pp. 23–25 (in Russian). https://doi.org/10.24411/2071-6168-2019-10207
6. Popova S.V., Khodyrev I.A. (2012) Keyword extraction. *Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics*, no. 1 (77), pp. 67–71 (in Russian).
7. Eliseeva E.N. (2019) Financial instruments for assessing the insolvency of industrial enterprises. *Region: systems, economy, management*, no. 3 (46), pp. 132–140 (in Russian).

8. Medvedev D.A. (2019) Big data: the reasons for their emergence and how they can be used. *Science and Education Today*, no. 4 (39), pp. 14–16 (in Russian).

9. Pang B., Lee L. (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135. https://doi.org/10.1561/1500000011

10. Morozov A.N. (2018) Alternative sources of statistical information as the basis for political decision making. *Problems of State and Municipal Management*, no. 2, pp. 50–70 (in Russian).

11. Puzanov A.S., Trutnev E.K., Markvart E., Popov R.A., Safarova M.D. (2017) *Strategic planning and urban regulation at the municipal level*. Moscow: Delo (in Russian).

12. Andreeva N.A., Ugrimova S.N. (2019) To the question of the application of statistical methods of integral estimation of effectiveness of system of management of industrial enterprises. *Accounting and Statistics*, no. 1 (53), pp. 42–49 (in Russian).

13. Norton D. P., Kaplan R.S. (2008). *Balanced scorecard*. Moscow: Olymp-Business.

14. Chugunov V.R., Zhukova L.V., Kovalchuk I.M., Kovaleva A.S. (2017) Mathematical methods of data grouping for making management decisions in planning tasks. *Actual Problems of System and Software Engineering 2017. Proceedings of the 5th International Conference on Actual Problems of System and Software Engineering Supported by Russian Foundation for Basic Research. Project #17-07-20565*, pp. 333–341 (in Russian).

15. Baresyagin A.A., Kupriyanov M.S., Stepanenko V.V., Kholod I.I. (2007) *Data analysis technologies: Data Mining, Visual Mining, Text Mining*, OLAP. 2nd edition. St. Petersburg: BXV-Petersburg (in Russian).

16. Shalymov D.S. (2008) Stable clustering algorithms based on index functions and stability functions. *Stochastic optimization in computer science*, vol. 4, pp. 236–248 (in Russian).

17. Kaufman L., Rousseeuw P. (2005) *Finding groups in data: An introduction to cluster analysis*. Wiley-Interscience.

18. Semina T.A. (2020) Sentiment analysis: modern approaches and existing problems. *Social sciences and humanities. Domestic and foreign literature. Ser. 6, Linguistics*, no. 4, pp. 47–64 (in Russian).

19. Liu B. (2010) Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition* (eds. N. Indurkhya, F.J. Damerau). London: Chapman and Hall/CRC.

20. Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinsky E.S., Lukashevich N.V., Sapin A.S. (2017) *Automatic text processing in natural language and data analysis. Tutorial*. Moscow: HSE (in Russian).

21. Demidova L.A., Stepanov M.A. (2019) An approach to solving problem of the structural transformations detection in the time series' groups. *Cloud of science*, no. 2. pp. 201–226 (in Russian).

22. Popova S.V., Khodyrev I.A. (2012) Extraction of keyword combinations. *Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics*, no. 1 (77), pp. 67–71 (in Russian).

23. Krasnyansky M.N., Obukhov A.D., Solomatina E.M., Voyakina A.A. (2018) Comparative analysis of machine learning methods for solving the problem of classifying documents of scientific and educational institution. *Bulletin of Voronezh State University*, no. 3, pp. 173–182 (in Russian).

24. Zhukova L.V. (2021) Express-analysis of the state of industrial enterprises of Moscow using the universal comprehensive indicator. *Economic Science of Modern Russia*, vol. 4 (95), pp. 89–96 (in Russian).

About the authors

Tatiana K. Bogdanova

Cand. Sci. (Econ.);

Assistant Professor, Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: tanbog@hse.ru

ORCID: 0000-0002-0018-2946

Liudmila V. Zhukova

Assistant Professor, Department of Applied Economics, Faculty of Economic Sciences, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: lvzhukova@hse.ru

ORCID: 0000-0003-1647-5337