Янин Д. М., Барковский С.С.

ТЕХНОЛОГИЯ КЛАССИФИКАЦИИ И ВЫЯВЛЕНИЯ ДУБЛИРОВАНИЯ РЕЗУЛЬТАТОВ ИНТЕЛЛЕКТУАЛЬНОЙ ДЕЯТЕЛЬНОСТИ В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ

Аннотация: *Предметом исследования является процедура учета неопределенности при выявлении сходства и дублирования между результатами интеллектуальной деятельности, обусловленной наличием семантической и прагматической неоднозначности в естественно-языковой форме описания результатов. Целью исследования является обеспечение идентификации дублированных результатов интеллектуальной деятельности в автоматизированном режиме, в котором эксперты только управляют ее параметрами путем определения состава классификаторов используемых при построении отношения сходства, упрощение межведомственной координации учета результатами интеллектуальной деятельности и целостности соответствующей базы данных, а также повышение достоверности и оперативности процедуры выявления дублированных результатов интеллектуальной деятельности за счет комбинирования частотно-семантических методов и формализованных знаний экспертов. Методология исследований основана на системном анализе, распознавании образов, синтезе баз данных и баз знаний, автоматической обработке текстов на естественном языке. В результате исследования разработан подход к формированию баз знаний классов информационных образов результатов интеллектуальной деятельности позволит классифицировать их различные типы и систематизировать знания о классах результатов интеллектуальной деятельности, за счет использования оригинального способа автоматизированной классификации и выявления дублирования результатов интеллектуальной деятельности.*
Ключевые слова: *результаты интеллектуальной деятельности, естественный язык, база данных, координация исследований, экспертная информация, классификация научных результатов, семантическая неоднозначность, прагматическая неоднозначность, описание научных результатов, исключение дублирования информации*
Review: *The article reviews a procedure of solving the uncertainty in identifying the similarities and duplications between the results of intellectual activity, due to the presence of semantic and pragmatic ambiguity in natural language form of results description. The aim of the study is to provide identification of duplicate results of intellectual activity in an automated mode, in which experts only manage its settings by determining the composition of the classifiers used in the construction of relations of similarity, simplification of inter-agency coordination based on the results of intellectual activity and integrity of the database, as well as increasing the reliability and*

DOI: 10.7256/2305-6061.2015.2.16683

efficiency of the procedure identifying duplicate the results of intellectual activity by combining the frequency-semantic methods and formalized expert knowledge. The research methodology is based on a system analysis, pattern recognition, synthesis databases and knowledge bases, the automatic processing of natural language texts. The authors developed an approach for building knowledge bases classes for information representations of the results of intellectual activity that would classify its different types and systematize knowledge about the classes of intellectual property through the use of the author's method of automated classification and identification of duplications of intellectual property.

Keywords: *pragmatic ambiguity, semantic ambiguity, classification of scientific results, expert information, research coordination, database, natural language, results of intellectual activity, description of the scientific results, avoiding duplication of information*

Результаты анализа состояния и направлений развития отечественных и зарубежных систем учета и контроля результатов интеллектуальной деятельности (РИД) свидетельствуют о том, что в настоящее время в нашей стране и за рубежом развитию и совершенствованию систем учета и контроля использования РИД, как в рамках государства, так и в рамках отдельных ведомств и организаций уделяется значительное внимание [1, 2]. В целом, работы по совершенствованию направлены на комплексную автоматизацию процесса учета и контроля использования РИД с использованием перспективных информационных технологий, позволяющих объединить информационные ресурсы отдельных субъектов, выполняющих функции учета и контроля использования РИД в рамках корпоративных автоматизированных информационных систем [3, 4].

При развитии методического обеспечения системы учета и контроля использования РИД важнейшим направлением является совершенствование способов классификации и поиска РИД, выявления необоснованного дублирования. Анализ существующих способов [5, 6] решения этой задачи с учетом характеристик информационной среды и объекта автоматизации показал, что в соответствии с условиями идентификации данных о РИД реализацию автоматизированного способа классификации и выявления дублирования РИД рационально производить одновременно на основе нескольких наиболее адекватных подходах. Частотно-семантический метод дает высокую скорость оценки, но недостаточную достоверность и точность оценок [7, 8]. Экспертные методы дают точность и качество оценок, однако требуют большего времени обработки и предъявляют высокие требования к квалификации лица, принимающего решение [9]. Комбинирование статистических и экспертных методов удовлетворяет требованиям по оперативности, но обеспечивает снижение достоверности (по отношению к экспертному) за счет сложности объектов анализа (РИД) и информационного представления их семантического и прагматического содержания [10, 11]. Повышение качества классификации и выявления дублирования РИД с сохранением оперативности частотно-семантического метода возможно за счет:

– построения дискретной модели предметной области в виде базы знаний по классам РИД, что позволит обобщить и структурировать знания о проводимых

DOI: 10.7256/2305-6061.2015.2.16683

разработках и получаемых результатах;

– использования знаний экспертов не при непосредственном сравнении и выявления необоснованного дублирования РИД, а при заведении данных по ним, т.е. когда происходит их позиционирование в многомерном пространстве признаков (рубрикаторе) и определение класса РИД в созданной модели предметной области;
– усовершенствования автоматизированной технологии формализации данных о РИД в интересах увеличения количества полезной информации, повышения ее полноты и структурирования.

Решение подобных задач идентификации дублирования объектов (РИД), классификации, методическое обеспечение которых опирается на несколько формальных теорий (а в данном случае преобладают принципы и методы теории распознавания образов и теории информации), предусматривает общую стандартную последовательность операций: определение системы координат; выбор и формирование метрики для измерения в системе координат меры близости между объектами; построение алгоритма обработки первичных параметров, характеризующих объекты анализа в многомерном пространстве признаков (систем координат), и отображения значений этих параметров посредством выбранной метрики в отношение сходства, а затем и подобия между объектами анализа [12-14].

Все множество существующих подходов к решению подобных задач выстраиваются по приведенной последовательности операций, но на разной методической и математической основе, определяемых конкретной решаемой задачей и характеристиками самих объектов анализа [15, 16]. Для решения задачи классификации и выявления дублирования РИД в автоматизированной системе управления развитием ВВСТ целесообразно использовать информационно-поисковую подсистему (рисунок 1), в которой в качестве базовых подходов рационально использовать следующие [17, 18]:

1. на основе метода идентификации подобия информационных образов (ИО) РИД с использованием комбинированного классификационно-дескрипторного языка представления знаний, объектно-ориентированного подхода и метрики близости ИО РИД в многомерном признаковом пространстве (модель №1);
2. на основе частотно-семантического метода (модель №2).

Методическое обеспечение данной подсистемы, основывается на способе формализации информационных образов РИД и формирования БЗ классов, который позволяет устранять трудности [18], связанные с естественно-языковым (ЕЯ) представлением данных по РИД и невозможностью их обработки как единого информационного объекта во внутреннем представлении автоматизированной системы.

Реализация способа включает этапы:

1. Предварительный лингвистический анализ;
2. Формализация дискретной модели внутреннего представления ИО РИД;
3. Формирование базы знаний по классам ИО.

DOI: 10.7256/2305-6061.2015.2.16683

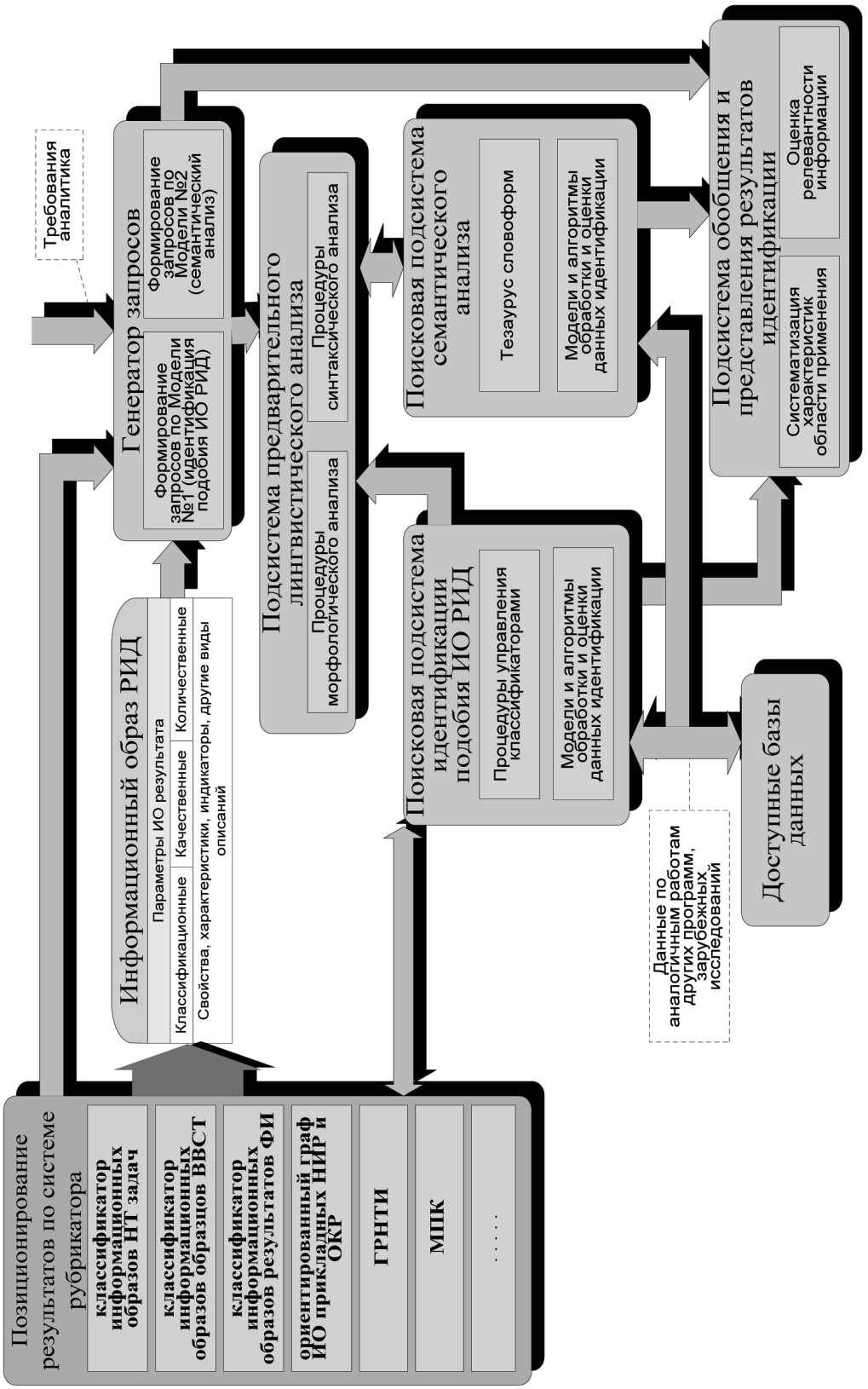

Рисунок 1 Структурная модель автоматизированной информационно-поисковой подсистемы классификации и выявления дублирования РИД

DOI: 10.7256/2305-6061.2015.2.16683

1. Предварительный лингвистический анализ данных по РИД, включает - морфологический и синтаксический анализ. Цель морфологического анализа состоит в получении основ (под основой понимается словоформа с отсеченным окончанием) со значениями грамматических категорий (например, часть речи, род, число, падеж) для каждой из словоформ. Используем точный и приближенный метод морфологического анализа. Точный метод базируются на использовании словаря основ слов или словоформ, приближенный - на экспериментально установленной связи между конечными буквосочетаниями словоформ и их грамматической информацией. В основе приближенных методов морфологического анализа лежит гипотеза, согласно которой по конечным буквам и буквосочетаниям можно практически однозначно определить грамматический класс слова.

В результате проведения морфологического анализа могут возникать неоднозначности при определении грамматической информации, которые снимаются после проведения синтаксического анализа. Задачей синтаксического анализа является осуществление грамматического разбора предложений на основе информации, заложенной в словаре. Любые средства синтаксического анализа состоят из двух частей: базы знаний о конкретном языке и собственно алгоритма синтаксического анализа, т.е. набора стандартных операторов, обрабатывающих текст на основе этих знаний. Источником знаний (грамматических) являются данные, полученные в результате морфологического анализа, а также различные таблицы, которые априорно заполнены стандартным образом и представляют собой результаты эмпирической обработки текстов с целью выделения определенных закономерностей, необходимых для проведения синтаксического анализа [10, 11, 16].

2. Формализация дискретной модели внутреннего представления ИО РИД. В качестве языка внутреннего представления данных применим модель, основанную на объектно-ориентированном подходе и комбинированном языке построения знаний (ЯПЗ), объединяющий в себе табличный язык и дескрипторный. Под ЯПЗ понимается специализированный искусственный язык, предназначенный для описания основного смыслового содержания поступающих в систему сообщений (данных по РИД), с целью обеспечения возможности последующего их поиска и обработки [2, 18].

2.1. Применение объектно-ориентированного подхода к формированию ИО РИД обеспечивает формирования информационных образов РИД (ИО РИД) и позволяет обращаться с поступающими данными, как с единым объектом (информационным образом), т.е. самодостаточным элементом, который в чем-то идентичен другим таким же однородным объектам в рамках класса, но в то же время отличается от них определенными уникальными свойствами или те же свойства обладают иным качественным и количественным значением.

2.2. Формирование комбинированного информационно-поискового языка. Так как представление данных о РИД осуществляется в естественно-языковом формате, являющимся универсальной знаковой системой, служащей для обмена информацией, необходимо учесть его недостатки, основным из которых является многообразие средств передачи смысла. Кроме лексики естественного языка функцию передачи смысла выполняет ряд других элементов [1]: контекст; парадигматические отношения между словами;

DOI: 10.7256/2305-6061.2015.2.16683

-текстуальные отношения между словами; ссылки на слова (словосочетания, сокращения, фразы и т.д.), ранее упоминавшиеся в тексте сообщения.

ЕЯ формат представления данных по таким сложным объектам анализа как РИД обеспечивает наличие значительной семантической и прагматической неопределенности. Решением данной проблемы является использование для внутреннего представления данных при построении модели предметной области (базы знаний классов ИО РИД) искусственного языка представления знаний (ЯПЗ), который создается на базе ЕЯ, однако отличается от него компактностью, наличием четких грамматических правил и значительным снижением семантической неопределенности.

При выборе ЯПЗ необходимо учесть:

– неоднородность источников данных об объектах анализа;
– возможность предварительно устанавливать отношения между терминами и понятиями предметной области, т.е. формирование основы и настройке базы знаний экспертами;
– способность к структурным изменениям и увеличению знаний (обучению).

Таким требованиям удовлетворяет комбинированный язык построения знаний, объединяющий в себе табличный язык и дескрипторный [12]. Табличный язык позволяет снизить влияние неоднородности объектов анализа – данных по РИД, обусловленной разнообразием источников данных. Так как все РИД имеют определенный стандартный набор признаков, то применение одноформатного ввода данных в соответствии с шаблоном позволяет упростить процедуру формирования информационных образов РИД. Необходимо использовать шаблон, включающий как стандартную форму РИД, так и дополнительную расширенную информацию, полезную для проведения анализа свойств РИД.

Дескрипторный язык обеспечит устранение ограничений по построению семантических образов. Объединенная модель представляет собой информационную дескрипторную систему [12. 16], которая может быть изображена в виде иерархического графа верхними уровнями, которого являются позиции шаблона. Каждая позиция соответствует либо иерархическому классификатору, либо линейному. В вершинах графа размещаются дескрипторы, а дуги соответствуют связям этих дескрипторов. По существу дескрипторный язык представляет собой ограниченную дискретную модель профессионального языка, который тоже дискретен и может в свою очередь рассматриваться как ограниченная модель естественного языка.

3. Формирование базы знаний по классам ИО. С использованием полученной дискретной модели внутреннего представления ИО РИД осуществляется формирование базы знаний классов ИО РИД. В целом система формирования БЗ классов ИО РИД основывается на элементах:

– объектно-ориентированный подход;
– комбинированный ЯПЗ (шаблон +рубрикатор);
– модель представления знаний в виде сети фреймов.

Названные элементы взаимосвязаны (перекрывают друг друга) и относятся к различным уровням [10, 11, 12, 16] описания (исследования) системы знаний о классах ИО

DOI: 10.7256/2305-6061.2015.2.16683

РИД. Объектно-ориентированный подход относится к концептуальному, комбинированный ЯПЗ к структурному, модель представления знаний в виде сети фреймов, как объединяющий элемент, к функциональному уровню. Анализ систем построения знаний показал, что фреймовая модель представления знаний позволяет объединить объектно-ориентированный подход и комбинированный ЯПЗ, а так же обладает рядом достоинств других моделей знаний: декларативных, процедурных, семантических сетей и моделей, использующих алгебру нечетких множеств.

Особенностью данной модели является введение в систему представления знаний модульной структуры в виде фреймов, которые представляют собой синтаксически-семантические блоки в общем случае процедурально-декларативного типа. Класс и соответственно ИО РИД является реализацией модели знаний о РИД в виде сети фреймов. Фреймы представляют собой локальные семантические сети [12] и являются единицами информации соответствующие позициям шаблона. В целом заполненная база знаний представляет собой иерархическую сеть классов, полнота и целостность которой поддерживается и отслеживается экспертами. База знаний классов ИО РИД относится к открытым системам, в которую, по мере необходимости, можно добавлять новые компоненты (классы, механизмы ЯПЗ).

База знаний обладает большой гибкостью за счет объединения объектно-ориентированного подхода, комбинированного ЯПЗ (шаблон +рубрикатор) и модели представления знаний в виде сети фреймов, что обеспечивает существование определенной структуры предметной области (рубрикатора), описывающей изначальный синтаксис, но при этом обеспечивается возможность:

- формирования новых частных понятий;
- установления и изменения правил-отношений между ними, не заложенными в синтаксис;
- конструирование фреймов, классов и ИО РИД различного типа и принадлежности к существующим научно-техническим разработкам;
- настройки БЗ на различные изменения в рубрикаторах и типах РИД.

Применение этого подхода к формированию БЗ классов ИО РИД позволит классифицировать различные типы РИД и систематизировать знания о классах РИД. Сравнение РИД в такой форме затруднительно, поэтому процедура оценки сходства РИД должна опираться на предложенный способ автоматизированной классификации и выявления дублирования РИД, основанный на формализации информационных образов РИД и формирования БЗ классов, позволяет:

- обеспечить идентификацию дублированных РИД в автоматизированном режиме, в котором эксперты только управляют ее параметрами путем определения состава классификаторов используемых при построении отношения сходства, что обеспечивает его адаптивное определение;
- облегчить операции по межведомственной координации РИД и обеспечить целостность данных в виде ИО РИД, что дает возможность при обработке данных оперировать с ними как с единым объектом;

DOI: 10.7256/2305-6061.2015.2.16683

– повысить достоверность и оперативность процедуры выявления дублированных РИД за счет комбинирования частотно-семантических методов и формализованных знаний экспертов.

Библиография :

1. Буренок В.М., Ляпунов В.М., Мудров В.И. Теория вооружения (учебное пособие) / Под ред. А.А.Рахманова. М.: 46 ЦНИИ МО РФ, 2002. 88 с.
2. Барковский С.С., Желтов П.В., Лукашов А.М. Подход к формализации модели семантической структуры текста в системах документооборота // Вестник Казанского государственного технического университета им. А.Н. Туполева. 2010. № 2. С. 96-100.
3. Голосовский М.С. Модель жизненного цикла разработки программного обеспечения в рамках научно-исследовательских работ // Автоматизация. Современные технологии. 2014. № 1. С. 43-46.
4. Голосовский М.С. Информационно-логическая модель процесса разработки программного обеспечения // Программные системы и вычислительные методы. 2015. № 1. С. 59-68.
5. Богомолов А.В. Методика формирования индекса состояния объекта по результатам многомерной статистической классификации // Информационные технологии. 2000. № 12. С. 45.
6. Шипилов В.В., Куксин К.Г., Баранов Н.А. Управление ресурсами при обеспечении безопасности защищаемых объектов // Нелинейный мир. 2014. Т. 12. № 7. С. 29-32.
7. Щеглов И.Н., Богомолов А.В., Печатнов Ю.А. Исследование влияния репрезентативности обучающей выборки на качество работы методов распознавания образов // Нейрокомпьютеры: разработка, применение. 2002. № 9-10.
8. Кукушкин Ю.А., Бухтияров И.В., Богомолов А.В. Обобщение результатов независимых экспериментальных исследований методом мета-анализа // Информационные технологии. 2001. № 6. С. 48.
9. Козлов В.Е., Богомолов А.В., Рудаков С.В., Оленченко В.Т. Математическое обеспечение обработки рейтинговой информации в задачах экспертного оценивания // Мир измерений. 2012. № 9. С. 42-49.
10. Максимов И.Б., Столяр В.П., Богомолов А.В. Прикладная теория информационного обеспечения медико-биологических исследований. Москва: Бином, 2013. 311 с .
11. Кукушкин Ю.А., Богомолов А.В., Ушаков И.Б. Математическое обеспечение оценивания состояния материальных систем // Информационные технологии. 2004. № 7 (приложение). 32 с.
12. Кузин Л.Т. Основы кибернетики. Т.2. Основы кибернетических моделей.-М.: Энергия, 1973. 148 с.
13. Фёдоров М.В., Калинин К.М., Богомолов А.В., Стецюк А.Н. Математическая модель автоматизированного контроля выполнения мероприятий в органах военного управления // Информационно-измерительные и управляющие системы. 2011. Т. 9. № 5. С. 46-54.
14. Барковский С.С., Воробьёв А.А. Технология планирования ресурсного обеспечения федеральных целевых программ // Финансы и управление. 2015. № 3. С. 11-24.
15. Щеглов И.Н., Богомолов А.В., Печатнов Ю.А. Алгоритм формирования обучающей выборки искусственной нейронной сети // Нейрокомпьютеры: разработка, применение. 2000. № 2. С. 12.

DOI: 10.7256/2305-6061.2015.2.16683

16. Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В.. Базы данных. Интеллектуальная обработка информации. М.: «Нолидж», 2000. 144 с.
17. Шибанов Г.П. Современные технологии проведения обликовых исследований // Автоматизация. Современные технологии. 2015. № 9. С. 26-33.
18. Надежность и эффективность в технике: Справочник: В 10 т. Том 3. Эффективность технических систем. / Под общ. ред. В.Ф. Уткина-М.: Машиностроение, 1988. 232 с.

References:

1. Burenok V.M., Lyapunov V.M., Mudrov V.I. Teoriya vooruzheniya (uchebnoe posobie) / Pod red. A.A.Rakhmanova. M.: 46 TsNII MO RF, 2002. 88 s.
2. Barkovskii S.S., Zheltov P.V., Lukashov A.M. Podkhod k formalizatsii modeli semanticheskoi struktury teksta v sistemakh dokumentooborota // Vestnik Kazanskogo gosudarstvennogo tekhnicheskogo universiteta im. A.N. Tupoleva. 2010. № 2. S. 96-100.
3. Golosovskii M.S. Model' zhiznennogo tsikla razrabotki programmnogo obespecheniya v ramkakh nauchno-issledovatel'skikh rabot // Avtomatizatsiya. Sovremennye tekhnologii. 2014. № 1. S. 43-46.
4. Golosovskii M.S. Informatsionno-logicheskaya model' protsessa razrabotki programmnogo obespecheniya // Programmnye sistemy i vychislitel'nye metody. 2015. № 1. S. 59-68.
5. Bogomolov A.V. Metodika formirovaniya indeksa sostoyaniya ob"ekta po rezul'tatam mnogomernoi statisticheskoi klassifikatsii // Informatsionnye tekhnologii. 2000. № 12. S. 45.
6. Shipilov V.V., Kuksin K.G., Baranov N.A. Upravlenie resursami pri obespechenii bezopasnosti zashchishchaemykh ob"ektov // Nelineinyi mir. 2014. T. 12. № 7. S. 29-32.
7. Shcheglov I.N., Bogomolov A.V., Pechatnov Yu.A. Issledovanie vliyaniya reprezentativnosti obuchayushchei vyborki na kachestvo raboty metodov raspoznavaniya obrazov // Neirokomp'yutery: razrabotka, primenenie. 2002. № 9-10.
8. Kukushkin Yu.A., Bukhtiyarov I.V., Bogomolov A.V. Obobshchenie rezul'tatov nezavisimykh eksperimental'nykh issledovanii metodom meta-analiza // Informatsionnye tekhnologii. 2001. № 6. S. 48.
9. Kozlov V.E., Bogomolov A.V., Rudakov S.V., Olenchenko V.T. Matematicheskoe obespechenie obrabotki reitingovoi informatsii v zadachakh ekspertnogo otsenivaniya // Mir izmerenii. 2012. № 9. S. 42-49.
10. Maksimov I.B., Stolyar V.P., Bogomolov A.V. Prikladnaya teoriya informatsionnogo obespecheniya mediko-biologicheskikh issledovanii. Moskva: Binom, 2013. 311 s.
11. Kukushkin Yu.A., Bogomolov A.V., Ushakov I.B. Matematicheskoe obespechenie otsenivaniya sostoyaniya material'nykh sistem // Informatsionnye tekhnologii. 2004. № 7 (prilozhenie). 32 s.
12. Kuzin L.T. Osnovy kibernetiki. T.2. Osnovy kiberneticheskikh modelei.-M.: Energiya, 1973. 148 s.
13. Fedorov M.V., Kalinin K.M., Bogomolov A.V., Stetsyuk A.N. Matematicheskaya model' avtomatizirovannogo kontrolya vypolneniya meropriyatii v organakh voennogo upravleniya // Informatsionno-izmeritel'nye i upravlyayushchie sistemy. 2011. T. 9. № 5. S. 46-54.
14. Barkovskii S.S., Vorob'ev A.A. Tekhnologiya planirovaniya resursnogo obespecheniya federal'nykh tselevykh programm // Finansy i upravlenie. 2015. № 3. S. 11-24.

DOI: 10.7256/2305-6061.2015.2.16683

15. Shcheglov I.N., Bogomolov A.V., Pechatnov Yu.A. Algoritm formirovaniya obuchayushchei vyborki iskusstvennoi neironnoi seti // Neirokomp'yutery: razrabotka, primenenie. 2000. № 2. S. 12.
16. Korneev V.V., Gareev A.F., Vasyutin S.V., Raikh V.V.. Bazy dannykh. Intellektual'naya obrabotka informatsii. M.: «Nolidzh», 2000. 144 s.
17. Shibanov G.P. Sovremennye tekhnologii provedeniya oblikovykh issledovanii // Avtomatizatsiya. Sovremennye tekhnologii. 2015. № 9. S. 26-33.
18. Nadezhnost' i effektivnost' v tekhnike: Spravochnik: V 10 t. Tom 3. Effektivnost' tekhnicheskikh sistem. / Pod obshch. red. V.F. Utkina-M.: Mashinostroenie, 1988. 232 s.

DOI: 10.7256/2305-6061.2015.2.16683