

---

## КОМПЬЮТЕРНАЯ ГРАФИКА И ВИЗУАЛИЗАЦИЯ

---

УДК 004.92

### АЛГОРИТМИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ОБРАБОТКИ МНОГОМЕРНЫХ ГЕТЕРОГЕННЫХ ДАННЫХ

© 2019 г. А. А. Захарова<sup>a,\*</sup>, С. Г. Небаба<sup>b,\*\*</sup>, Д. А. Завьялов<sup>c,\*\*\*</sup>

<sup>a</sup> Брянский государственный технический университет,  
241035 Брянск, бульвар 50 лет Октября, д. 7, Россия

<sup>b</sup> Национальный исследовательский Томский политехнический университет  
634050 Томск, пр. Ленина, д. 30, Россия

<sup>c</sup> Институт прикладной математики им. М.В. Келдыша  
125047 Москва, Миусская пл., д. 4, Россия

\* e-mail: zaa@tpu.ru

\*\* e-mail: stepan-lfx@tpu.ru

\*\*\* e-mail: zda@tpu.ru

Поступила в редакцию 16.02.2019 г.

После доработки 17.03.2019 г.

Принята к публикации 17.03.2019 г.

В работе предложен алгоритм выбора методов обработки многомерных гетерогенных данных на основании общих свойств используемых данных и методов, включенных в рассмотрение. Алгоритм реализован в виде программного обеспечения и проведено сравнение группы алгоритмов интерполяции на примере задачи построения 3D модели лица человека для системы распознавания личности. Показано, что предложенный алгоритм выбора методов обработки данных успешно работает для группы методов интерполяции данных.

DOI: 10.1134/S0132347419040101

#### 1. ВВЕДЕНИЕ

В настоящее время активно развиваются компьютерное зрение, робототехника, экспериментальная физика и многие другие научные направления. Задачи, возникающие в этих областях знаний, все больше связаны с обработкой больших массивов разнородных данных, то есть данных, полученных от датчиков и сенсоров разного типа, разрешающей способности, точности и частоты измерений. Датчиком или сенсором в контексте этого определения является любой источник дискретной цифровой информации, а конкретная совокупность синхронизированных между собой датчиков формирует мультисенсорную систему. Значимость проблемы обработки и анализа данных такого типа постоянно растет, при этом традиционные подходы к обработке данных теряют актуальность из-за растущих требований к комплексности и интеллектуальности решений [1, 2].

Быстрый и правильный выбор подходящего метода обработки разнородных (гетерогенных) данных имеет большое значение, и алгоритм, способный автоматически сделать такой выбор, может стать основной частью интеллектуальных

систем, работающих с многомерными массивами гетерогенных данных. Вопрос надежности и точности подобных систем остается ключевым, особенно для таких критичных направлений как автономные транспортные средства, системы безопасности, медицинские технологии, моделирование и прогнозирование природных и социальных событий [3]. Это делает задачу создания алгоритма автоматического выбора методов обработки многомерных гетерогенных данных для конкретных случаев актуальной и важной.

В данной работе предпринимается попытка создания алгоритма анализа и классификации методов обработки данных, который бы позволял выбрать наиболее эффективный способ для поставленной задачи, оперируя исключительно общими характеристиками массивов данных, такими как тип данных, размерность, объем выборок, область применения и ожидаемые результаты обработки.

#### 2. КЛАССИФИКАЦИЯ ВИДОВ И ТИПОВ ДАННЫХ

Одним из важных этапов алгоритма анализа и классификации методов должен быть первона-

чальный анализ типов и видов данных, присутствующих в массиве.

Существуют различные классификации видов и типов данных, эти классификации выбираются в зависимости от области получения и использования данных. Наиболее распространенной классификацией типов данных в области информатики и программирования является разбиение данных на:

- логические;
- числовые (целочисленные, с плавающей запятой);
- строковые;
- комплексные и прочие типы [4].

Виды данных, в свою очередь, описываются такими категориями, как:

- числовые данные;
- текстовые данные;
- мультимедийные данные (графические, звуковые) [5].

От вида и типа данных, которые необходимо обрабатывать, зависит точность результата и вычислительная сложность алгоритмов, нацеленных на получение конечного искомого результата. Ряд методов обработки неприменим к тем или иным типам или видам данных, либо применение не имеет смысла в контексте решаемой задачи. Следовательно, крайне важна выработка оптимального подхода к группировке методов по типам и видам обрабатываемых данных, что позволит не только структурировать существующие подходы к обработке данных, но и обнаружить новые зависимости и возможности применения классических методов и алгоритмов.

### 3. МЕТОДЫ, СПОСОБЫ И АЛГОРИТМЫ ОБРАБОТКИ МНОГОМЕРНЫХ ГЕТЕРОГЕННЫХ ДАННЫХ

Методы и способы обработки многомерных гетерогенных данных можно разделить по нескольким критериям: по предметной области получения данных (например: данные сейсморазведки, данные химического анализа, визуальные данные и т.п.), по принципу обработки (например: снижение размерности данных, вычисление интегральных и дифференциальных характеристик, корреляционный анализ, визуализация и когнитивный анализ [6] и т.п.). В общем случае, метод обработки выбирается исходя из предметной области, поставленных задач и объема данных, так как различные методы обладают различной вычислительной сложностью и точностью.

Кроме того, многие данные могут иметь различный способ представления, для данных большой размерности, описывающих процессы во времени, имеет смысл представление в частотной, фазовой или временной области. Для этих

видов данных существуют специфические способы обработки, например, преобразование Фурье, фильтрация полосовыми фильтрами, фильтрами высоких и низких частот.

В качестве объекта анализа рассматривались группы методов корреляционного анализа, регрессионного анализа, интерполяции и экстраполяции, методы интегрального и дифференциального анализа.

Интерполяция данных подразумевает повышение плотности данных за счет новых синтезированных промежуточных элементов. Вопрос точности новых данных решается с помощью экспериментальных данных в конкретной предметной области с большой частотой дискретизации, позволяющей провести сравнение натурных значений с вычисленными в ходе интерполяции.

Экстраполяция данных представляет собой прогнозирование и дополнение данных за пределами измеряемого интервала. Точность также может быть оценена по натурным экспериментальным наблюдениям.

Дифференцирование и интегрирование оценивают скорость изменения данных и общий кумулятивный эффект выбранных данных.

### 4. АЛГОРИТМ КЛАССИФИКАЦИИ И ВЫБОРА КОНКРЕТНЫХ СПОСОБОВ ОБРАБОТКИ И АНАЛИЗА ДАННЫХ

Ключевой идеей предлагаемого подхода является отнесение данных к конкретной группе методов анализа и обработки по их свойствам с последующим выбором конкретных методов. Общая схема подхода представлена на рисунке 1 и может быть выражена следующим алгоритмом:

а) данные, которые необходимо обработать, оцениваются по следующим критериям: размерность (число исследуемых параметров), объем, типы и виды данных, решаемая задача и форма представления результата %.

б) исходя из свойств данных, определенных на предыдущем этапе, выбирается группа методов обработки данных. В частности, если для числовых данных спектр существующих методов достаточно широк, то для текстовых и мультимедийных данных набор методов чаще всего кардинально отличается, и требует либо декомпозиции данных, либо их предварительного анализа и преобразования в иной вид представления;

с) внутри выбранной группы методов почти у всех вариаций существуют свои собственные ограничения или условия применимости, такие как размерность обрабатываемых данных, объем выборки данных, чувствительность к характеру изменения данных, быстродействие или вычислительная сложность. В зависимости от ранее учтенных свойств данных на этапе выбора кон-

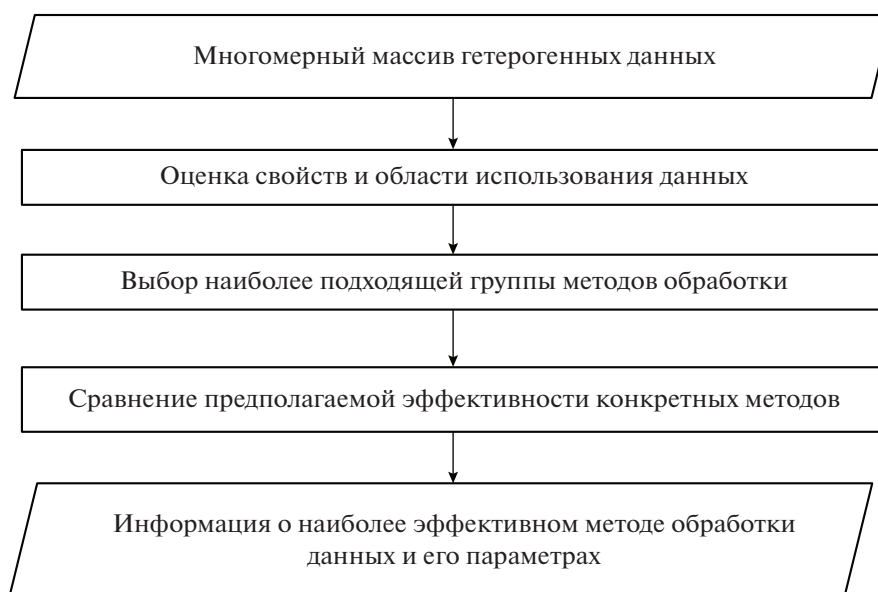


Рис. 1. Общая схема предлагаемого алгоритма.

кретного метода необходимо сравнить существующие методы по тем их свойствам, которые являются объективными и априорно известными;

д) в результате выбора конкретного метода обработки и анализа данных необходимо получить список его свойств и параметров, подходящих для рассматриваемой задачи и дающих приемлемый результат.

В работе подробно рассматривается этап сравнения предполагаемой эффективности конкретных методов из группы по общим свойствам и характеристикам данных и методов обработки.

В рамках данной работы было решено остановиться на методах интерполяции данных, как на распространенном и гибком инструменте обработки числовых данных различной природы и размерности.

Известно, что интерполяция – это способ нахождения промежуточных значений величины по имеющемуся дискретному набору известных значений [7].

Многим из тех, кто сталкивается с научными и инженерными расчетами, часто приходится оперировать наборами значений, полученных опытным путем или методом случайной выборки. Как правило, на основании этих наборов требуется построить функцию, на которую могли бы с высокой точностью попадать другие получаемые значения. Такая задача называется аппроксимацией. Интерполяцией называют такую разновидность аппроксимации, при которой кривая построенной функции проходит точно через имеющиеся точки данных.

Существует также близкая к интерполяции задача, которая заключается в аппроксимации какой-либо сложной функции другой, более простой функцией. Если некоторая функция слишком сложна для производительных вычислений, можно попытаться вычислить ее значение в нескольких точках, а по ним построить, то есть интерполировать, более простую функцию. Разумеется, использование упрощенной функции не позволяет получить такие же точные результаты, какие давала бы первоначальная функция. Но в некоторых классах задач достигнутый прирост скорости вычислений может перевесить получаемую погрешность в результатах.

К настоящему времени существует множество различных способов интерполяции. Выбор наиболее подходящего алгоритма зависит от ответов на вопросы: как точен выбираемый метод, каковы затраты на его использование, насколько гладкой является интерполяционная функция, какой количества точек данных она требует и т. п.

Наиболее известные способы интерполяции: интерполяция методом ближайшего соседа, интерполяция многочленами, линейная интерполяция, интерполяционная формула Ньютона, метод конечных разностей, ИМН-1 и ИМН-2, многочлен Лагранжа (интерполяционный многочлен), схема Эйткена, сплайн-функция, кубический сплайн, рациональная интерполяция, тригонометрическая интерполяция.

Обратное интерполирование (вычисление  $x$  при заданной  $y$ ) осуществляется также различными способами: полином Лагранжа, обратное интерполирование по формуле Ньютона, обратное интерполирование по формуле Гаусса, интерполя-

ция функции нескольких переменных, билинейная интерполяция, бикубическая интерполяция.

Также нельзя не упомянуть в контексте оптимизации обработки гетерогенных данных проблему энтропии данных.

Энтропия данных определяет меру необратимого распределения (увеличения) данных или бесполезности (избыточности) данных, так как не все полученные данные (метаданные) извлекаемые из анализируемой мультисенсорной среды можно использовать для превращения в какую-нибудь полезную работу по принятию решений.

С этой точки зрения предлагаемый алгоритм может быть в перспективе дополнен функцией предварительной оценки энтропии данных и исключения избыточной информации из дальнейшей обработки, таким образом, позволяя снизить вычислительную стоимость всех последующих операций [8].

## 5. СОЗДАНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ АЛГОРИТМА КЛАССИФИКАЦИИ И ВЫБОРА МЕТОДОВ И ПРОВЕРКА ЕГО ЭФФЕКТИВНОСТИ

Реализация алгоритма выбора и применения метода анализа данных для конкретных задач имеет большое значение для апробации общей концепции классификации методов обработки и демонстрации повышения эффективности методов по точности и скорости получения результата в различных областях. В качестве примеров апробации предлагаемого алгоритма могут быть выбраны области компьютерного зрения (стереофотоъемка с помощью ИК-датчика глубины и обработка предварительно рассчитанных данных), безопасность предприятий (комплексный анализ данных от множества датчиков, фиксирующих критически важные показатели на производстве: влажность, давление, температура, состав атмосферы) и другие.

Так как этап сравнения конкретных методов предлагаемого алгоритма анализа и классификации проверяется на группе методов интерполяции данных, которые имеют широкое применение в обработке цифровых многомерных данных, то для апробации предлагаемого алгоритма решено использовать данные задачи компьютерного зрения. Задача обработки и сравнения 3D-моделей лиц связана с проблемой повышения плотности точек 3D-модели интерполированием известных с целью увеличения качества 3D-модели.

**Задача обработки и сравнения 3D-моделей.** Данные при построении 3D-моделей могут быть представлены в различном виде, однако чаще всего используется массив данных, включающий в себя координаты каждой из точек модели в пространстве, а также текстуру модели, которая име-

ет информацию о том, какая точка текстуры соответствует точке в пространстве. Для объектов простых форм или для объектов, которые описаны очень подробно, могут применяться и другие подходы хранения данных, однако для них чаще всего не применяется метод интерполяции.

В работе [9] в качестве способа совмещения фотографии лица человека и 3D-модели использовалась интерполяция модифицированным методом Шепарда, которая была подобрана с помощью подробного литературного обзора и обзора случаев, в которых применяются те или иные методы. В рамках предлагаемого алгоритма анализа и классификации предполагается, что аналогичный выбор можно сделать, оперируя исключительно формальной информацией о данных и методах.

В качестве характеристик, на основе которых сравниваются алгоритмы интерполяции, были выбраны следующие:

- Число регулируемых параметров. Каждый алгоритм, помимо исходных данных, может требовать подбора одного или нескольких числовых параметров, которые влияют на точность и даже корректность получаемого результата. Такие параметры могут иметь диапазон допустимых и целесообразных значений в зависимости от других свойств данных и алгоритма. Очевидно, что чем меньше у алгоритма регулируемых параметров и чем более формально описан процесс их подбора, тем проще и быстрее автоматизировать подбор и применение алгоритма обработки данных.

- Вычислительная сложность. Существует общепризнанный способ сравнения вычислительной сложности алгоритмов, который показывает порядок времени работы алгоритма — оценка временной сложности алгоритма. Оценка отображается как функция  $T(n)$  (или  $O(n)$ ), где природа функции показывает вычислительную сложность, таким образом, время работы некоторых алгоритмов выражается как  $O(1) < O(\log n) < O(n) < O(n^2) < O(n!)$ .

- Размерность данных, при которой алгоритм может быть применен или достигает наибольшей эффективности. Многие алгоритмы обработки данных либо не могут дать корректный результат на исходных данных большой размерности, либо их эффективность существенно снижается.

- Минимальный и максимальный объем массивов данных, на которых может быть получен результат. Некоторые алгоритмы теряют эффективность либо работоспособность, если объем исходных данных слишком мал или слишком велик, это важно учитывать на этапе подбора методов обработки.

- Чувствительность к ошибкам данных, выбросам, удаленным значениям. Чрезмерная чувствительность, свойственная некоторым алгоритмам, может привести к некорректному результату при

**Таблица 1.** Ключевые характеристики алгоритмов интерполяции

| Название метода                                                       | Число регулируемых параметров | Вычислительная сложность | Дополнительные особенности                                                                                                          |
|-----------------------------------------------------------------------|-------------------------------|--------------------------|-------------------------------------------------------------------------------------------------------------------------------------|
| Модифицированный метод Шепарда [10]                                   | 2                             | $O(\log N)$              | Наибольшее быстродействие при размерности данных 2–5, на 100 и более элементах массивов                                             |
| Обратные взвешенные расстояния (Inverse distance weighting, IDW) [10] | нет                           | $O(N)$                   | Чувствительность к удаленным значениям и выбросам в данных                                                                          |
| Метод ближайшего соседа [11]                                          | нет                           | $O(N)$                   | Наиболее грубый метод, плохо подходит для нерегулярных сеток                                                                        |
| Бикубическая интерполяция [12]                                        | 2                             | $O(N)$                   | Хорошие возможности по параллельным вычислениям, поддержка больших объемов данных, возможность использования на нерегулярных сетках |
| Радиальные базисные функции (Radial basis functions, RBF) [13]        | 1                             | $O(N \cdot \log(N))$     | Одинаково высокая эффективность почти на любых размерностях и объемах данных                                                        |

недостаточном анализе общих свойств данных и алгоритмов.

В таблице 1 представлены характеристики ряда алгоритмов, которые были выбраны в качестве основных для группы методов интерполяции. Ключевыми характеристиками принято число регулируемых параметров и вычислительная сложность, остальные характеристики перечислены в столбце “Дополнительные особенности”.

Как можно видеть из таблицы 1, представленные алгоритмы интерполяции существенно отличаются по точности результата, вычислительной сложности и количеству параметров, которые необходимо подбирать непосредственно для конкретной задачи. От выбора метода будет существенно зависеть точность решения задачи, скорость получения этого решения, его общая корректность и соответствие поставленной цели.

В качестве критерия выбора метода обработки данных для предлагаемого алгоритма, помимо очевидных ограничений по размерности и объему данных, рационально ввести весовые коэффициенты значимости, отражающие важность того или иного параметра для конкретной задачи: точность результата, время вычисления, сложность подготовки и настройки регулируемых параметров.

В рассматриваемой задаче построения 3D-моделей финальным критерием качества интерполяции может выступать как точность системы распознавания объекта по интерполированным данным, так и близость интерполированной модели к реальным параметрам объекта. Ранее [9] сравнение проводилось с использованием системы распознавания личности по изображению лица человека, и конечная точность интерполяции могла быть оценена исходя из точности определе-

ния личности с использованием интерполяции и без нее.

В данной работе проведено прямое сравнение на конкретной 3D-модели лица человека, полученной методом стереосъемки и различными методами интерполяции эталонных 3D-моделей лиц. На рисунке 2 представлен пример растрового фото, стереоизображения и интерполированной 3D-модели лица, на которых проводилось сравнение. Характеристики данных, на которых было проведено сравнение, следующие:

- Размер растрового изображения лица: 400\*600 точек.

- Число точек эталонных 3D-моделей, по которым строится интерполяция: 25782.

- Число точек 3D-модели лица, полученной по стереоизображению: 24046.

Все массивы данных были нормированы по общим ключевым точкам, данные представлены в трехмерном пространстве, информация о текстуре и цвете лица была отброшена, так как она не использовалась в дальнейшей обработке данных (в соответствии с идеей снижения энтропии данных).

Методы, требующие задания параметров интерполяции, использовались с рекомендованными параметрами для заданных условий.

В качестве оценки точности интерполяции взято среднее евклидово расстояние всех точек 3D-моделей, полученных методом стереосъемки и методом интерполяции эталонных 3D-моделей. В качестве оценки скорости интерполяции взято время вычисления всех интерполированных точек синтетической 3D-модели. Для реализации методов интерполяции и их сравнения была написана программа на языке C++ в среде разработки Visual Studio.

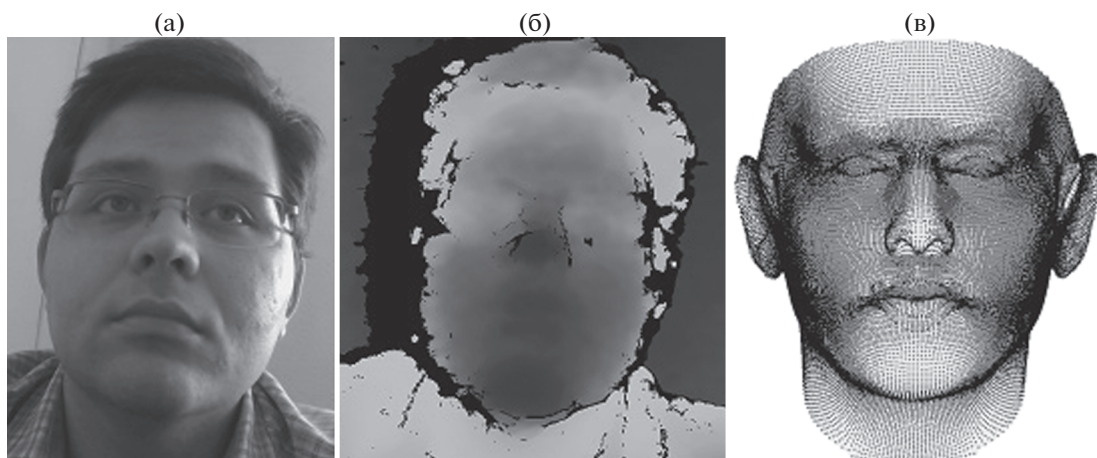


Рис. 2. а) Растровое изображение лица; б) стереоизображение; в) 3D-модель.

Результаты сравнения алгоритмов представлены в таблице 2.

Как можно видеть из таблицы 2, наименьшей погрешности интерполяции удалось добиться при использовании метода радиальных базисных функций, хотя по скорости вычисления он существенно уступает другим методам, в том числе и изначально использовавшемуся модифицированному методу Шепарда.

Из проведенного сравнения по таблицам 1 и 2 можно сделать вывод о том, что большинство методов интерполяции могут быть классифицированы по вычислительной сложности и числу регулируемых параметров. Первый критерий влияет как на точность получаемого результата, так и на скорость обработки данных, второй критерий серьезно влияет на время обработки из-за необходимости многократного перебора параметров интерполяции, кроме того, наличие или отсутствие параметров регулировки напрямую влияет на сложность автоматизации обработки данных.

Таким образом, дальнейшая работа по оценке критериев эффективности методов обработки данных и созданию алгоритма автоматизированной классификации и выбора оптимальных методов представляется достаточно актуальной.

## 6. ВЫВОДЫ

В работе проведен обзор видов и типов данных, методов и алгоритмов обработки. Предложен алгоритм выбора конкретных способов обработки данных на основании общих свойств данных и рассматриваемых алгоритмов.

Проведено сравнение группы алгоритмов интерполяции на примере задачи построение 3D-модели лица человека для системы распознавания личности.

Анализ результатов показал, что предложенный алгоритм выбора методов обработки данных успешно работает для группы методов интерполяции данных, выявив на основании общих характеристик методов более точный метод, чем подобранный вручную на основании литературного обзора, а также достаточно правильно оценив порядок быстродействия всех методов интерполяции. По результатам выбора метода радиальных базисных функций достигнута наименьшая погрешность построения 3D-модели и наибольшее время вычисления, как и предсказано на основе классификации, лежащей в основе предлагаемого алгоритма. Этот подход представляется целесообразным автоматизировать и в перспективе расширить на более общую задачу выбора групп методов анализа данных.

Таблица 2. Результаты сравнения эффективности алгоритмов

| Название метода                                                  | Средняя погрешность интерполяции | Время вычисления, с |
|------------------------------------------------------------------|----------------------------------|---------------------|
| Модифицированный метод Шепарда                                   | 1.4                              | <u>6.57</u>         |
| Обратные взвешенные расстояния (Inverse distance weighting, IDW) | 2.2                              | 11.42               |
| Метод ближайшего соседа                                          | 12.6                             | 7.02                |
| Бикубическая интерполяция                                        | 2.8                              | 10.81               |
| Радиальные базисные функции (Radial basis functions, RBF)        | <u>1.2</u>                       | 16.75               |

## БЛАГОДАРНОСТИ

Исследование выполнено за счет гранта Российского научного фонда, проект 18-11-00215.

## СПИСОК ЛИТЕРАТУРЫ

1. Клеменков П.А., Кузнецов С.Д. Большие данные: современные подходы к хранению и обработке // Труды ИСП РАН. Т. 23. 2012. С. 143–158.
2. *Budylskii D., Podvesovskii A.* (et al.) Hierarchical Deep Learning: A Promising Technique for Opinion Monitoring and Sentiment Analysis in Russian-Language Social Networks // A. Kravets et al. (Eds.): CIT&DS 2015, Communications in Computer and Information Science, V. 535 – Springer International Publishing. 2015. P. 583–592.
3. Истомин Е.П., Колбина О.Н., Степанов С.Ю., Сидоренко А.Ю. Анализ моделей и систем обработки гетерогенных данных для использования в прикладных ГИС // Научный вестник. 2015. № 4. С. 53.
4. Пирс Б. Типы в языках программирования. Добросвет, 2012. 680 с. ISBN 978-5-7913-0082-9.
5. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. СПб.: Питер, 2017. 336 с.
6. *Zakharova A.A., Vekhter E.V., Shklyar A.E., Pak A.Y.* Visual modeling in an analysis of multidimensional data // Journal of Physics: Conference Series. IOP Publishing. 2018. V. 944. № 1.
7. Берг Й., Лёфстрём Й. Интерполяционные пространства. Введение. М.: Мир, 1980. 264 с.
8. Габидулин Э.М., Пилипчук Н.И. Лекции по теории информации. МФТИ, 2007. 214 с.
9. Небаба С.Г., Захарова А.А. Алгоритм построения деформируемых 3D-моделей человеческого лица и обоснование его применимости в системах распознавания личности // Труды СПИИРАН. 2017. Вып. 52. С. 157–179.
10. *Shepard D.* A two-dimensional interpolation function for irregularly-spaced data // Proceedings of the 1968 23rd ACM national conference. ACM. 1968. P. 517–524.
11. *Bovik A.C.* Handbook of image and video processing. Academic press, 2010. 1384 p.
12. *Keys R.* Cubic convolution interpolation for digital image processing // IEEE transactions on acoustics, speech, and signal processing. 1981. V. 29. № 6. P. 1153–1160.
13. *Loeffler C. F., Zampogno L., Mansur W.J., Bulcão A.* Performance of Compact Radial Basis Functions in the Direct Interpolation Boundary Element Method for Solving Potential Problems // CMES – Computer Modeling in Engineering and Sciences. 2017. V. 113. P. 387–412.