21. *Kerschberg L., Kim W., Scime A.* Personalizable semantic taxonomy-based search agent. USA: George Mason Intellectual Properties, INC (Fairfax, VA), 2006.
22. *Kerschberg L., Jeong H., Kim W.* Emergent Semantic in Knowledge Sifter: An Evolutionary Search Agent based on Semantic Web Services, *Journal on Data Semantics*. VI. LNCS. Vol. 4090. Springer, Heidelberg, 2006, pp. 187-209.

Статью рекомендовал к опубликованию д.т.н., профессор М.М. Ошхунов.

Кравченко Юрий Алексеевич – Южный федеральный университет; e-mail: yakravchenko@sfedu.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634371651; кафедра систем автоматизированного проектирования; доцент.

Kravchenko Yury Alekseevich – Southern Federal University; e-mail: yakravchenko@sfedu.ru; 44, Nekrasovskiy lane, Taganrog, 347928, Russia; phone: +78634371651; the department of computer aided design; associate professor.

УДК 002.53 DOI 10.18522/2311-3103-2016-7-1828

С.Б. Картиев, В.М. Курейчик

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АЛГОРИТМА РЕШЕНИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ ДЛЯ ОСУЩЕСТВЛЕНИЯ ВОПРОСНО-ОТВЕТНОГО ПОИСКА В ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЕ ПРОГНОЗИРОВАНИЯ*

Данная статья посвящена проблемам построения модуля вопросно-ответного поиска неструктурированной информации в информационно-аналитической системе прогнозирования относительно коллекции исходных состояний сложной технической системы. Современные информационно-поисковые системы основаны на принципах поиска по ключевым словам. Данный вид поиска предоставляет на выходе коллекцию веб-страниц, которая по вероятности может содержать нужный материал для пользователя. В статье предлагается подход приведения задачи кластеризации к оптимизационной задаче и ее решения с использованием метаэвристических методов. Дано введение в традиционные методы кластеризации, приведены их преимущества и недостатки. Кластеризация является частным случаем обучением без учителя. Отсутствие учителя предусматривает то что в системе нет эксперта, который может присваивать документам классы. Приведено описание основной модели вычислений и средств хранения данных, применяемых в разработанной системе. Предложен подход к построению подобных модулей и их математическое обеспечение, которое является решением некоторых проблем обработки естественного языка. Новизна работы заключается в использовании модифицированного генетического алгоритма для решения задачи кластеризации текстовых документов, который позволяет параллельно анализировать ряд наилучших решений.. Это позволяет повысить качество подсистемы поиска информационно-аналитической системы (ИАС) прогнозирования. Подсистема поиска ИАС применяется для извлечения информации для прогнозирования из коллекции исходных состояний сложной технической системе. Произведена разработка модифицированного генетического алгоритма кластеризации. Приведена программная реализация модуля информационного поиска ИАС прогнозирования с использованием разработанного алгоритма на языке Java для решения задачи кластеризации и применение библиотеки OpenNLP для обработки естественного языка. Также определено место разработанного модуля в системе диагностирования сложных технических систем по поддержанию работоспособности программной системы. Проведены тестовые испытания подобной системы на последней версии копии сайта Wikipedia.org. Эксперименты показали уменьшение времени выполнения алгоритма и улучшение качества полученных результатов.

Генетический алгоритм; кластеризация; информационный поис; прогнозирование.

* Работа выполнена за счет полного финансирования по ГЗ 01/ОПНИ и гранта РФФИ № 15-07-05523.

S.B. Kartiev, V.M. Kureychick

DESIGN AND ANALYSIS OF ALGORITHM FOR SOLVING CLUSTERING PROBLEM FOR QUESTION-ANSWERING MODULE OF FORECASTING SYSTEM

This article is dedicated to problems of construction of the module for question-answer search unstructured information in the information-analytical prediction system relative to the collection of the initial states of complex technical systems. Modern information and search engines are based on the principles of search by keywords. This type of search provides the output of web pages collection that probability may contain the necessary material to the user. The paper proposes an approach to the task of bringing the clustering optimization problem and solve it using metaheuristic methods. An introduction to traditional clustering methods, given their advantages and disadvantages. Clustering is a special case of learning without a teacher. The lack of teachers is provided that the system has no expert who can assign the document class. The description of the basic model of computing and data storage means used in the developed system. An approach to the construction of such modules and their software, which is the solution of some problems of natural language processing. The novelty of the work lies in the use of the modified genetic algorithm for solving the problem of clustering of text documents that allows you to simultaneously analyze a number of the best solutions .. This allows you to improve the quality of the search subsystems of information-analytical system (IAS) prediction. IAS search subsystem is used to retrieve information for the prediction of the collection of the initial states of complex technical system. Produced development of modified genetic clustering algorithm. Shows the software implementation of information retrieval module IAS forecasting algorithm developed using the Java programming language to solve the problem of clustering and application OpenNLP libraries for natural language processing. Also defines the place of the developed module in the system diagnostics of complex technical systems to maintain the health of a software system. The testing of such a system to copy the latest version Wikipedia.org site. The experiments showed a decrease execution time of the algorithm and improve the quality of the results.

Genetic algorithm; clustering; information retrieval; prediction.

Введение. В условиях развития современного информационного общества, все более актуальной становится проблема структуризации информации, которая поступает из различных источников. Главным источником информации на данный момент является интернет. Количество информации в интернете практически никто не берется измерить. На примерах поиска информации с использованием поисковых систем можно сделать определенный вывод о том, что количество информации и информация, которая необходима для разрешения потребностей -> 1: 100000. Современные поисковые системы основаны на принципах поиска по ключевым словам. Данный вид поиска предоставляет на выходе коллекцию вебстраниц, которая по вероятности может содержать нужный материал для пользователя. В работе предполагается подход приведения задачи кластеризации к оптимизационной задаче и ее решения с использованием метаэвристических методов. Принципиальным отличием работы является использование генетических алгоритмов, позволяющих параллельно анализировать ряд наилучших решений.

1. Классификация систем вопросно-ответного поиска. Целью работы является улучшения результатов поиска конченого состояния технической системы.

Опишем основные этапы осуществления информационного поиска:

1. Индексирование.
2. Ввод запроса пользователем.
3. Ранжирование.
4. Выдача результатов.

Первый этап, называемый индексированием, подразумевает под собой сбор коллекции исходных данных и их индексирование для дальнейшей поддержки функционирования информационно-поисковой системы (ИПС), Данный процесс

производится постоянно для поддержания актуальности выборки. Второй этап - на экран пользователя выводится окно ввода запроса, которое из себя представляет пустое поле и кнопку “Поиск”, ранжирование – ИПС для каждого найденного элемента вычисляют его степень соответствия смыслу запроса пользователя. После проведения этих действий происходит этап выдачи результатов поиска.

Основной проблемой современных поисковых систем является формализация вопроса и распознавание его смысла и выделение из контекста. Решением данной проблемы является подход, называемый вопросно-ответным поиском, который распознает вопрос, заданный пользователем и выдает короткий и лаконичный ответ, сформулированный на естественном языке [1]. Первые исследования в данной области были произведены в 1961 году для индексации и выделения логического смысла выражения, которое задано вычислительной машине при помощи перфокарт [2]. Авторы считают, что для решения задачи определения контекста в подсистеме ИАС прогнозирования возможно использование систем вопросно-ответного поиска.

Рассмотрим классификацию вопросно-ответных систем (ВОС).

На современном этапе развития, типы вопросов, которые могут быть заданы системе разбиваются на две категории:

1. Вопросы на фактическое определение.
2. Повествовательные вопросы.

Основными парадигмами вопросно-ответного поиска являются:

1. Системы, основанные на принципах информационного поиска.
2. Системы, основанные на знаниях.

Приведем алгоритм работы ВОС, основанный на принципах информационного поиска:

Алгоритм IRQA (Вопросно-ответный информационный поиск).

Вход: Вопрос, заданный пользователем.

Выход: Краткий ответ заданный на естественном языке.

1. Формулирование вопроса и определение типа вопроса.
2. Информационный поиск (предварительно должна производиться процедура индексации).
3. Получение от этапа 2 коллекции релевантных документов.
4. Ранжирование.
5. Получение ранжированных документов
6. Обработка ответа.
7. Выдача ответа на естественном языке.

Первым этапом ВОИС является работа модуля анализа вопросов. На вход подается вопрос, заданный на естественном языке. Для заданного вопроса выделяется его фокус, опора и семантическая метка.

- Фокус вопроса – сведения, которые несут в себе контекст вопроса, который пользователь передал для ожидания ответа.
- Опора вопроса – часть вопроса, которая содержит в себе информацию, которая подтверждает выбор конкретного ответа.
- Семантическая метка вопроса – класс запрашиваемой пользователем информации. [3]

2) Постановка задачи кластеризации текстовых документов. Кластеризация является разделением совокупности документов на подмножества, называемыми кластерами. Кластеры должны являться однородными изнутри, но обязаны четко отличаться от документов другого кластера [4]. Кластеризация является частным случаем обучением без учителя. Отсутствие учителя предусматривает то что в системе нет эксперта, который может присваивать документам классы. Раз-

ница между кластеризацией и классификацией изначально может показаться незначительной. Однако классификация является разновидностью обучения с учителем. Основная цель классификации – воспроизведение данных по категориям, которые установлены экспертом. В задаче кластеризации эксперта нет.

Сформулируем постановку задачи кластеризации.

Дано множество документов D = $\{d_1, d_2, \dots, d_k\}$, где k – количество кластеров, которое может задаваться и d ≠ ∅. Также существует целевая функция, которая оценивает качество кластеризации.

$D \rightarrow \{1, \dots, K\}$ – min (max) целевой функции

Кластеризация применяется для численного определения значения близости между документами

Реализация постановки задачи заключается в разработке генетического алгоритма для решения задачи кластеризации.

Основной количественной мерой в задаче кластеризации является метрика [4]. В качестве метрики используется расстояние между точками на плоскости, в большинстве случаев используется евклидово расстояние. Метрика является важнейшим инструментом, при помощи которого измеряют качество кластеризации. Кластеризация разделяется на два типа: плоская и иерархическая. Плоская кластеризация – совокупность кластеров, не имеющих явных взаимосвязей, а иерархическая кластеризация создает иерархию кластеров. Иерархические алгоритмы сначала помещают каждую точку в отдельный кластер. Далее ближайшие кластеры объединяются с использованием критерия близости. Данный процесс прекращается, когда дальнейшее объединение приводит к нежелательным кластерам, которые могут привести лишь к некачественному результату кластеризации.

Алгоритмы плоской кластеризации основаны на отнесении точек. Точки рассматриваются в определенном порядке, и каждая относится к наиболее подходящему ей кластеру. В качестве предварительных действий алгоритма происходит оценка начальных кластеров.

Кластеризация применяется для следующих задач:

1. Кластеризация результатов поиска
2. Разбиение и объединение подмножества массивов данных.
3. Кластеризация массивов данных
4. Языковые модели
5. Кластерный поиск.

Для численного определения значения близости между документами в задаче кластеризации используется метрика Минковского [5], в которой в качестве базиса используются многомерные евклидовые пространства.

$$M_p(\vec{x}, \vec{y}) = (\sum_{k=1}^{N} (x_k - y_k)^2$$

Однако, многомерные евклидовые пространства обладают рядом интуитивно неочевидных свойств, которые иногда называют “проклятием размерности” (ПР). Эти аномалии характерны и для неевклидовых пространств. Одно из проявлений “проклятия” – тот факт, что при большом числе измерений почти все пары точек находятся на одинаковом расстоянии друг от друга. Другие проявления – почти любые два вектора почти ортогональны [5].

При проявлении ПР появляются следующие проблемы:

- Трудоемкость вычислений.
- Необходимость хранения большого количества данных.
- Увеличение количества шумов в данных.
- Проявляются проблемы переобучения и мультиколлинеарности.

Одним из основных способом решения ПР является понижение размерности пространства при помощи проецирования на подпространство меньшей размерности [6]. Опишем основные методы кластеризации текстовых документов.

3. Обзор методов кластеризации текстовых документов.

3.1. Матричный латентно-семантический анализ. Латентно семантическая индексация (LSA) основана на сингулярном разложении матриц, где массиву документов ставится в соответствие матрица, строки которой соответствуют документам, а столбцы размером словаря. Метод LSA широко применяется при ранжировании выдачи информационно-поисковых систем, основанных на цитировании. Это алгоритм HITS (Hyperlink Induced Topic Search) – один из двух самых известных в области информационного поиска. Метод LSA не нуждается в предварительной настройке на специфический набор документов, вместе с тем позволяет качественно выявлять скрытые факторы. К недостаткам метода можно отнести невысокую производительность. Скорость вычисления SVD соответствует порядку $O(N2 \cdot k)$, где $N = |D| + |T|$, D – множество документов, T – множество термов, k – размерность пространства факторов. LSA также не предусматривает возможность пересечения кластеров, что противоречит практике. Кроме того, ввиду своей вычислительной трудоемкости метод LSA применяется только для относительно небольших матриц.

Достоинства:

- LSA не нуждается в предварительной настройке.
- Позволяет качественно определять скрытые факторы в выборке.

Недостатки:

- Пересечение кластеров не предполагается в LSA.
- Матрицы должны быть малого размера.
- Невысокая производительность

3.2. Метод k-средних. Основным методом кластеризации является алгоритм к-средних. В основе данного алгоритма лежит проход по всем элементом выборки, в которой каждая точка, кроме первоначально выбранных К точек, относятся к ближайшему кластеру:

- Инициализация. Пользователь выбирает число кластеров и назначает им гипотетические центры.
- Обновление кластеров. Каждый объект приписывается одному из центров по правилу минимального расстояния. Те объекты, которые приписаны центру образуют кластер. В качестве расстояния используется квадрат Евклидова расстояния.
- Обновление центров. Вычисляется центр масс каждого кластера, который и назначается новым центром.
- Правило остановки. Новые центры сравниваются с предыдущими. Если они совпадают, то конец работы алгоритма [7].

Достоинства:

- Вычислительная сложность O(kn)

Недостатки:

- Каждый объект кластеризации может попасть лишь в один кластер.

3.3. Метод суффиксных деревьев. Изначально метод суффиксных деревьев (Suffix Tree Clustering) был разработан для быстрого поиска подстрок в строках. Суффикс W строки S — это такая строка, в которой конкатенация (сцепление строк) VW совпадает с S для некоторой (возможно, пустой) строки V. Суффикс называется собственным, если $|V| \neq 0$. Суффиксное дерево – это дерево, содержащее все суффиксы строки [8]. Оно состоит из вершин, ветвей и суффиксных указателей. Метка узла в дереве определяется как конкатенация подстрок, маркирую-

щих ребра пути от корня дерева до этого узла. Существуют алгоритмы, реализующие построение суффиксных деревьев за $O(n)$ шагов, где n – длина строки. Ветви дерева обозначаются отдельными буквами или частями суффиксов строки [9].

Достоинства:

- Наглядность представления результатов.
- Высокая скорость работы.
- Кластеры могут пересекаться.

Недостатки:

- Предполагается, что дерево полностью должно быть загружено в оперативную память.
- Размер суффиксного дерева сильно превосходит входные данные.

4. Генетический алгоритм кластеризации. Авторами был разработан генетический алгоритм кластеризации текстовых документов для осуществления поиска по базе исходных состояний в информационно-аналитической системе прогнозирования для осуществления обработки информации и вывода статистических данных, предоставляемых для решения задачи прогнозирования. Генетические алгоритмы (ГА) – поисковые алгоритмы, основанные на механизмах натуральной селекции и натуральной генетики. Они реализуют "выживание сильнейших" среди рассмотренных структур, формируя и изменяя поисковый алгоритм на основе моделирования эволюции поиска [10]:

- Математически представим метрику в виде:

$$M\ (C_1 \ldots\ldots C_n) = \sum_{i=1}^{n} \sum_{x \in c_i} |x_j - y_i|,$$

где $n -$ размерность пространства, k – количество кластеров

Основной задачей работы генетического алгоритма является поиск наиболее подходящих центров *y* такие, что метрика будет минимизирована.

Целевая функция – присвоим каждую точку из популяции к одному из кластеров C_j с центром Z_j согласно формуле:

$$|x_i - y_j| < |x_i - y_p|, p = 1 \ldots\ldots K, p \neq j.$$

После кластеризации – центры кластеров, зашифрованные в хромосоме, заменяют средние точки соответствующего кластера. Новый центр кластера рассчитывается согласно формуле:

$$y_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j, i = 1 \ldots\ldots K.$$

В итоге целевая функция выглядит так: F = 1/M;

Рассмотрим подробнее основные этапы выполнения генетического алгоритма кластеризации:

1) Инициализация популяции. Задаем размер популяции P. Случайно выбираем точки из обучающей выборки и присваиваем им метки центров кластеров. Процесс повторяется P раз, где один центр – одна хромосома. Например, допустим у нас размерность N = 2, количество кластеров K = 4 то хромосома будет в следующем виде: (53, 75), (14, 16), (30, 32), (70, 75), где ген представляет из себя пару чисел, которые характеризуют точку центра кластера [11].
2) Селекция. Хромосома представляет из себя номер копий, которые пропорциональны их целевой функции в популяции В качестве стратегии селекции выбрана Колесо рулетки [12]
3) Кроссинговер. Случайный процесс, который комбинирует информацию между двумя родительскими хромосомами для генерации двух дочерних хромосом. В качестве оператора кроссинговера выбран Одноточечный кроссинговер [13].

4) Мутация. Каждая хромосома подвергается мутации согласно определенной заданной вероятностью M. Генерируем число δ с помощью равномерного распределения в диапазоне от 0 до 1. v возьмем как позиция гена в хромосоме [14].

$$v \pm 2 * \delta * \mathrm{v}, \qquad \mathrm{v} \neq 0$$
$$v \pm 2 * \delta, \qquad \mathrm{v} = 0$$
$$v \pm \delta * v.$$

Критерий остановки: Количество итераций.

Данный генетический алгоритм обеспечивает линейную сложность выполнения, что является наибольшим преимуществом по сравнению с основными методами кластеризации объектов.

5. Программная реализация алгоритма кластеризации для задачи вопросно-ответного поиска и экспериментальные исследования. Алгоритм кластеризации был реализован с использованием языка программирования Java. Кластеризатор реализован с учетом формирования онтологического представления предметной области [15] и встроен в модуль вопросно-ответного поиска разработанной информационно-аналитической системы прогнозирования. Модуль вопросно-ответного поиска отвечает за выдачу структурированной информации по состоянию исследуемой системы в конкретные временные периоды, заданные исходным запросом. Также модуль предлагает возможные интерпретации поискового запроса на основе анализа взаимосвязей понятий и распределяет результаты поиска по разным интерпретациям вводимого запроса [16].

Информационно-аналитическая система использует технологию разработки корпоративных информационных систем Spring и язык программирования Java. Приведем основные особенности технологии Spring.

Ядро Spring основано на принципе инверсии управления (Inversion of Control – IoC), при котором создание и управление зависимостями между компонентами становятся внешними.

Реализация DI в Spring основана на двух ключевых концепциях Java – компонентах JavaBean и интерфейсах. При использовании Spring в качестве поставщика DI гибкость определения конфигурации зависимостей внутри своих приложений разнообразными путями (т.е. внешне в XML-файлах, с помощью конфигурационных Java-классов Spring или посредством аннотаций Java в коде) [17].

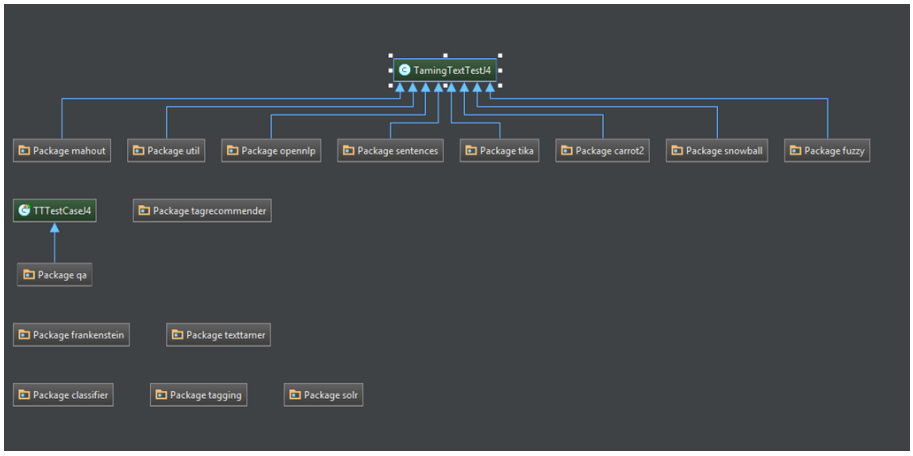

Рис. 1. Основная структура проекта QASystem

Задача прогнозирования представлена так: Дан временной ряд значений параметров исследуемого процесса [18]:

$$x_0, x_1 \dots \dots \dots . x_n; ,$$

где $x_i \in R$;

$$x_{t+d}(W) = f_t(x_1, \dots \dots, x_t; w)$$

Формула является моделью временного ряда, где *d= 1,.....,D*, *D* – горизонт прогнозирования. *W* – вектор параметров модели, x – единица статистического материала. Решение данной задачи состоит в том, что для прогнозирования на каждом шаге будет вырабатываться обучающая выборка, которая будет представлять из себя предыдущие состояния системы, которые влияют на N + 1 элемент временного ряда [19]. Разработанная ИАС прогнозирования использует подсистему вопросно-ответного поиска для построения возможных запросов оператора системы извлечения данных о техническом состоянии рассматриваемой системы. [20].

Эксперименты проводились с использованием компьютера Intel Core i5, 8 гб ОЗУ. В качестве исходных экспериментальных данных применялись 10,000 объектов кластеризации в двумерном пространстве, разделенном на 5 кластеров.

Размер популяции: 50

Количество генераций: 500

Время: 2000

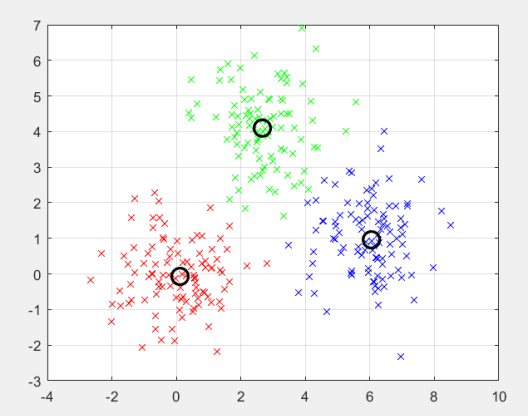

Рис. 1. Результат работы алгоритма

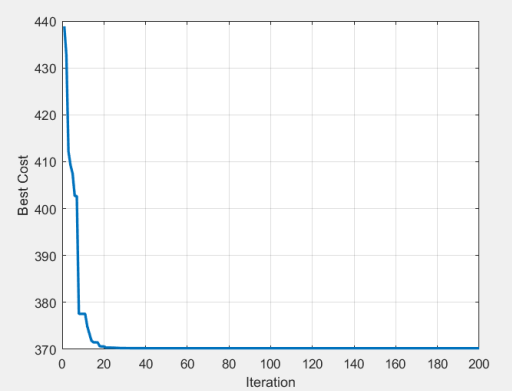

Рис. 2. Нахождение лучшего решения

Как показано на рисунке, генетический алгоритм кластеризации находит решение уже на ~20 итерации, что позволяет получить огромный выигрыш в производительности алгоритма поиска в ИАС прогнозирования для извлечения сведений о техническом состоянии исследуемой системы из общего количества данных документо-ориентированной СУБД.

Заключение. Работа была посвящена проблемам построения модуля вопросно-ответного поиска неструктурированной информации в информационно-аналитической системе прогнозирования относительно коллекции исходных состояний сложных технических систем. Был предложен подход к построению подобных модулей и их математическое обеспечение, которое является решением некоторых проблем обработки естественного языка. Новизна работы заключается в использовании генетического алгоритма для решения задачи кластеризации текстовых документов. Это позволяет повысить качество системы прогнозирования. Приведена программная реализация подобной системы с использованием разработанного алгоритма на языке Java и применением библиотеки OpenNLP. Проведены тестовые испытания подобной системы на последней версии копии сайта Wikipedia.org. Эксперименты показали уменьшение времени получения результатов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Simmons, Klein, McConlogue*. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196U204.
2. *Соловьёв А.А.* Синтаксические и семантические модели и алгоритмы в задаче вопросно-ответного поиска // Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.
3. *Соловьёв А.А., Пескова О.В.* Построение вопросно-ответной системы для русского языка: модуль анализа вопросов // Новые информационные технологии в автоматизированных системах: материалы 13-го научно-практического семинара. – М.: Моск. гос. ин-т электроники и математики. – 2010. – С. 41-49.
4. *Bishop C.* Pattern Recognition and Machine Learning, Springer, 2006.
5. *Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y.Wu.* An Efficient k-means Clustering Algorithm: Analysis and Implementation.
6. *Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman* Mining of Massive Datasets. – Cambridge University Press, 2014. – 511 p.
7. *Маннинг К., Рагхаван П., Шютце Х.* Введение в информационный поиск. – М.: Вильямс, 2011. – 528 с.
8. *Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman* Mining of Massive Datasets. – Cambridge University Press, 2014. – 511 p.
9. *Миркин Б.Г.* Введение в анализ данных: учебник для бакалавриата и магистратуры. – М.: Юрайт, 2014. – 174 с.
10. *Романовский И.В.* Дискретный анализ: учеб. пособие для студентов, специализирующихся по прикладной математике и информатике. – 4-е изд., испр. и доп. – СПб.: Невский Диалект; БХВ-Петербург, 2008. – 336 с.
11. *Ландэ Д.В, Снарский А.А.* Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Книжный дом «ЛИБРОКОМ», 2009. – 264 с.
12. *Емельянов В.В., Курейчик В.В., Курейчик В.М.* Теория и практика эволюционного моделирования. – М.: Физматлит, 2003. – 432 с.
13. *Гладков Л.А., Курейчик В.В., Курейчик В.М.* Генетические алгоритмы: учеб. пособие / под ред. В.М. Курейчика. – М.: Физматлит, 2004. – 400 с.
14. *Ломакина, Л.С., Губернаторов В.П.* Модификация эволюционно-генетического алгоритма для эффективного диагностирования сложных систем // Системы управления и информационные технологии. – 2013. – Т. 53, № 3. – С. 59-64.
15. *Наместников А.М.* Метауровень информационного обеспечения САПР: от теории к практике. – Ульяновск: УлГТУ, 2015. – 176 с.
16. *Гаврилова Т.А, Кудрявцев Д.В., Муромцев Д.И.* Инженерия знаний. Модели и методы: учебник. – СПб.: Изд-во Лань 2016. – 324 с.

17. *Кларенс Хо, Роб Харрон.* Spring 3 для профессионалов = Pro Spring 3. – М.: Вильямс, 2012. – 880 с.
18. *Картиев С.Б., Курейчик В.М.,* Алгоритм классификации, основанный на принципах случайного леса для решения задачи прогнозирования // Программные продукты и системы. – 2016. – № 2. – С. 11-15.
19. *Картиев С.Б., Курейчик В.М., Мартынов А.В.* Параллельный алгоритм прогнозирования коротких временных рядов // Труды Конгресса по интеллектуальным системам и информационным технологиям «IS&IT'15». Научное издание в 4-х т. – М.: Физматлит, 2015. – С. 27-47.
20. *Картиев С.Б., Курейчик В.М.,* Разработка распределенной системы анализа временных рядов на основе модели вычисления MapReduce // Труды Конгресса по интеллектуальным системам и информационным технологиям «IS&IT'16». Научное издание в 4-х т. – М.: Физматлит, 2016. – С. 36-43.

REFERENCES

1. *Simmons, Klein, McConlogue*. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196U204.
2. *Solov'ëv A.A.* Sintaksicheskie i semanticheskie modeli i algoritmy v zadache voprosno-otvetnogo poiska [Syntactic and semantic models and algorithms in the task of question-answering search], *Trudy 13-y Vserossiyskoy nauchnoy konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii» - RCDL'2011, Voronezh, Rossiya, 2011* [Proceedings of 13-th all-Russian scientific conference "digital libraries: advanced methods and technologies, digital collections" - RCDL'2011, Voronezh, Russia, 2011].
3. *Solov'ev A.A., Peskova O.V.* Postroenie voprosno-otvetnoy sistemy dlya russkogo yazyka: modul' analiza voprosov [Question-answering system Building for the Russian language: the module of analysis of issues], *Novye informatsionnye tekhnologii v avtomatizirovan-nykh sistemakh: materialy 13-go nauchno-prakticheskogo seminara* [New information technologies in automatedtion systems: proceedings of the 13th scientific-practical seminar]. Moscow: Mosk. gos. in-t elektroniki i matematiki 2010, pp. 41-49.
4. *Bishop C.* Pattern Recognition and Machine Learning, Springer, 2006.
5. *Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y.Wu.* An Efficient k-means Clustering Algorithm: Analysis and Implementation.
6. *Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman* Mining of Massive Datasets. Cambridge University Press, 2014, 511 p.
7. *Manning K., Ragkhavan P., Shyuttse Kh.* Vvedenie v informatsionnyy poisk [Introduction to information retrieval]. Moscow: Vil'yams, 2011, 528 p.
8. *Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman* Mining of Massive Datasets. Cambridge University Press, 2014, 511 p.
9. *Mirkin B.G.* Vvedenie v analiz dannykh: uchebnik dlya bakalavriata i magistratury [Introduction to data analysis: a textbook for undergraduate and graduate programs]. Moscow: Yurayt, 2014, 174 p.
10. *Romanovskiy I.V.* Diskretnyy analiz: ucheb. posobie dlya studentov, spetsializiruyushchikhsya po prikladnoy matematike i informatike [Discrete analysis: a textbook for students specializing in applied mathematics and computer science]. 4th ed. – SPb.: Nevskiy Dialekt; BKhV-Peterburg, 2008, 336 p.
11. *Lande D.V, Snarskiy A.A.* Internetika: Navigatsiya v slozhnykh setyakh: modeli i algoritmy [Internetika: Navigation in complex networks: models and algorithms]. Moscow: Knizhnyy dom «LIBROKOM», 2009, 264 p.
12. *Emel'yanov V.V., Kureychik V.V., Kureychik V.M.* Teoriya i praktika evolyutsionnogo modelirovaniya [Theory and practice of evolutionary modeling]. Moscow: Fizmatlit, 2003, 432 p.
13. *Gladkov L.A., Kureychik V.V., Kureychik V.M.* Geneticheskie algoritmy: ucheb. posobie [Genetic algorithms: a textbook], under the ed. of V.M. Kureychika. Moscow: Fizmatlit, 2004, 400 p.
14. *Lomakina, L.S., Gubernatorov V.P.* Modifikatsiya evolyutsionno-geneticheskogo algoritma dlya effektivnogo diagnostirovaniya slozhnykh sistem [Modification of evolutionary genetic algorithm for efficient diagnosis of complex systems], *Sistemy upravleniya i informatsionnye tekhnologii* [Control systems and information technology], 2013, T. 53, No. 3, pp. 59-64.
15. *Namestnikov A.M.* Metauroven' informatsionnogo obespecheniya SAPR: ot teorii k praktike [Meaurement information support of CAD: from theory to practice]. Ul'yanovsk: UlGTU, 2015, 176 p.
16. *Gavrilova T.A, Kudryavtsev D.V., Muromtsev D.I.* Inzheneriya znaniy. Modeli i metody: uchebnik [Knowledge engineering. Models and methods: textbook]. St. Petersburg: Izd-vo Lan' 2016, 324 p.

17. *Klarens Kho, Rob Kharrop.* Spring 3 dlya professionalov = Pro Spring 3 [Spring 3 for pros = Pro Spring 3]. Moscow: Vil'yams, 2012, 880 p.
18. *Kartiev S.B., Kureychik V.M.* Algoritm klassifikatsii, osnovannyĭ na printsipakh sluchaĭnogo lesa dlya resheniya zadachi prognozirovaniya [The classification algorithm is based on the principles of random forests for forecasting], *Programmnye produkty i sistemy* [Software Products and Systems], 2016, No. 2, pp. 11-15.
19. *Kartiev S.B., Kureychik V.M. Martynov A.V.* Parallel'nyy algoritm prognozirovaniya korotkikh vremennykh ryadov [A parallel algorithm for forecasting short time series], *Trudy Kongressa po intellektual'nym sistemam i in-formatsionnym tekhnologiyam «IS&IT'15». Nauchnoe izdanie v 4-kh t.* [Proceedings of Congress on intelligent systems and information technologies "IS&IT'15". Scientific publication in 4 vol.]. Moscow: Fizmatlit, 2015, pp. 27-47.
20. *Kartiev S.B., Kureychik V.M.* Razrabotka raspredelennoy sistemy analiza vremennykh ryadov na osnove modeli vychisleniya MapReduce [Development of a distributed system for analyzing time series based on the model of MapReduce computation], *Trudy Kongressa po intellektual'nym sistemam i informatsionnym tekhnologiyam «IS&IT'16». Nauchnoe izdanie v 4-kh t.* [Proceedings of Congress on intelligent systems and information technologies "IS&IT'16". Scientific publication in 4 vol.]. Moscow: Fizmatlit, 2016, pp. 36-43.

Статью рекомендовал к опубликованию д.т.н., профессор Я.Е. Ромм.

Картиев Санчир Басанович – Южный федеральный университет; e-mail: mlearningsystems@gmail.com; 347928, г. Таганрог, пер. Некрасовский, 44; кафедра ДМиМО; аспирант.

Курейчик Виктор Михайлович – e-mail: kur@tsure.ru; кафедра ДМиМО; профессор.

Kartiev Sanchir Basanovich – South Federal University; e-mail: mlearningsystems@gmail.com; 44, Nekrasovskiy, Taganrog, 347928, Russia; the department of discrete mathematics and optimization methods; postgraduate student.

Kureychick Viktor Michailovich – e-mail: kur@tsure.ru; the department of discrete mathematics and optimization methods; professor.

УДК 681.3.06:378.1 DOI 10.18522/2311-3103-2016-7-2839

И.И. Казмина, Е.В. Нужнов

МОДИФИКАЦИЯ АПРИОРНОГО АЛГОРИТМА ДЛЯ АНАЛИЗА ДАННЫХ УЧЕБНОГО ПРОЦЕССА ВУЗА*

Интеллектуальный анализ данных (ИАД) учебного процесса является одним из механизмов, позволяющих получить больше полезных сведений из имеющихся массивов данных и использовать полученные результаты с целью повышения эффективности и качества образовательной деятельности. Преимуществом использования ИАД является возможность выявления скрытых закономерностей, которые не всегда видны при использовании статистических методов. Априорный алгоритм ИАД позволяет на основе анализа больших массивов исходных данных выявлять зависимости между часто встречающимися элементами данных и анализируемой величиной. В качестве анализируемой величины в данном случае выступает успеваемость студентов, которая численно отражает эффективность учебного процесса, а исходными являются различные данные, касающиеся образовательной деятельности. Априорный алгоритм может выявить большое число правил в исходных данных, значительная часть которых может быть заранее известна пользователю, вследствие чего неинформативна. Для устранения этой проблемы в работе предлагается модификация Априорного алгоритма, учитывающая такой показатель правил, как их инфор-

* Исследование выполнено за счет гранта Российского научного фонда (проект № 14-11-00242) в Южном федеральном университете.