

## Технологии больших данных в электронном образовании

В последнее время электронное образование во всем мире стремительно развивается и основной проблемой становится своевременное обеспечение учащихся качественной учебной информацией. Эту задачу невозможно решить без анализа большого потока информации, поступающего в информационную среду электронного образования от участников образовательного процесса — студентов, преподавателей, администрации и т.д. В этой среде существует большое количество различных типов данных, как структурированных, так и неструктурированных, обработку которых трудно осуществить традиционными статистическими методами. Целью исследования является показать, что для разработки и внедрения успешных систем электронного обучения необходимо использовать новые технологии, которые позволили бы хранить и обрабатывать большие потоки данных.

Для хранения больших данных требуется большой объем дисковой памяти. Показано, что для решения этой проблемы эффективно использовать кластерную технологию NAS (Network Area Storage), позволяющая хранить информацию учебных заведений на NAS — серверах и иметь к ним общий доступ из Интернета. Для обработки и персонализации Больших Данных в среде электронного образования предлагается использовать технологии MapReduce, Hadoop, NoSQL и другие. В статье приводятся примеры использования этих технологий в облачной среде. Эти технологии в электронном образовании позволяют достигнуть гибкости, масштабируемости, доступности, качества обслуживания, безопасности, конфиденциальности и простоты использования учебной информации.

Другой важной проблемой электронного образования является выявление новых, порою скрытых, взаимосвязей в больших данных, новых знаний (data mining), которые могут быть использованы для улучшения образовательного процесса и повышения эффективности его управления. Для классификации электронных образовательных ресурсов, выявления паттернов (шаблонов) студентов со сходными психологическими, поведенческими и интеллектуальными характеристиками, разработки индивидуализированных учебных программ в статье предлагается использовать методы анализа больших данных.

В статье показано, что на сегодняшний день разработано множество программных приложений для интеллектуального анализа больших данных. Эти программные продукты можно использовать для классификации, кластеризации, регрессионного и сетевого анализа учебной информации. Применение этих методов в электронном образовании позволит педагогам своевременно получать информацию об обучающихся, оперативно реагировать на любые изменения процесса обучения, своевременно вносить изменения в учебный контент. Полученные результаты исследования предлагается использовать для выработки рекомендаций при создании электронных курсов в высших и средних учебных заведениях Азербайджана.

**Ключевые слова:** электронное образование, большие данные, анализ больших данных, персонализация данных, управление большими данными, MapReduce, Hadoop, NoSQL, data mining в электронном образовании.

Gyulara A. Mamedova, Lala A. Zeynalova, Rena T. Melikova

Azerbaijan National Academy of Sciences, Institute of Information Technology, Baku, Republic of Azerbaijan

## Big data technologies in e-learning

Recently, e-learning around the world is rapidly developing, and the main problem is to provide the students with quality educational information on time. This task cannot be solved without analyzing the large flow of information, entering the information environment of e-learning from participants in the educational process — students, lecturers, administration, etc. In this environment, there are a large number of different types of data, both structured and unstructured. Data processing is difficult to implement by traditional statistical methods. The aim of the study is to show that for the development and implementation of successful e-learning systems, it is necessary to use new technologies that would allow storing and processing large data streams.

In order to store the big data, a large amount of disk space is required. It is shown that to solve this problem it is efficient to use clustered NAS (Network Area Storage) technology, which allows storing information of educational institutions on NAS servers and sharing them with Internet. To process and personalize the Big Data in the environment of e-learning, it is proposed to use the technologies MapReduce, Hadoop, NoSQL and others. The article gives examples of the use of these technologies in the cloud environment. These technologies in e-learning allow achieving flexibility, scalability, availability, quality of service, security, confidentiality and ease of educational information use.

Another important problem of e-learning is the identification of new, sometimes hidden, interconnection in Big Data, new knowledge (data mining), which can be used to improve the educational process and improve its management. To classify electronic educational resources, identify patterns of students with similar psychological, behavioral and intellectual characteristics, developing individualized educational programs, it is proposed to use methods of analysis of Big Data.

The article shows that at present many software applications have been developed for the intellectual analysis of Big Data. These software products can be used for classification, clustering, regression and network analysis of training information. The application of these methods in e-learning will allow lecturers to receive timely information about students, promptly respond to any changes in the learning process, and timely make changes to educational content. The results of the research are proposed to be used to develop recommendations for the creation of electronic courses in higher and secondary educational institutions of Azerbaijan.

**Keywords:** e-learning, Big Data, Big Data Analysis, data personalization, managing Big Data, Hadoop, NoSQL, MapReduce, data mining in e-learning.

## Введение

В последнее время электронное обучение становится тенденцией изменения в системах образования многих стран мира. Так, в США электронное образование стало наиболее перспективной стратегией в системе национального образования. Об этом свидетельствуют данные американского консорциума электронного образования «Sloan» по которым, в общей сложности, осенью 2014 года 5,8 млн студентов обучались дистанционно, из них 2,85 млн. обучались on-line по всем курсам, а 2,97 миллиона — по некоторым курсам программы обучения [1].

В мировом масштабе на развитие электронного образования в 2011 году потрачено семь триллионов долларов и, по прогнозу аналитиков, каждый последующий год будет увеличиваться ежегодно на 25% [2].

В течение длительного периода в системах электронного образования учебных заведений накапливается огромное количество информации о различных аспектах образовательного процесса: студентах, их успеваемости и посещаемости, преподавателях и их научно-образовательной и административной деятельности, образовательного контента (текст, аудио, видео) и т.д. Эти данные необходимо эффективно хранить, обрабатывать и анализировать. Для обработки больших архивов и больших потоков данных требуются новые технологии, которые часто называют технологиями Big Data (Большие данные) [3–5].

Термин Big Data относится к большим и сложным наборам данных, которые могут быть структурированными, или неструктурированными и занимают очень большой объем дисковой памяти. В сфере электронного образования Большие данные охватывают три аспекта: **объем, скорость и разнообразие**.

Большой объем данных означает информацию о большом количестве обучающихся и тысячах учебных заведений. Эти данные, накапливаясь, дают море информации, которая может быть использована для эффективного управления учебным процессом [6].

Скорость изменения больших данных позволяет в интерактивном режиме контролировать процесс обучения и своевременно реагировать на любые изменения учебного процесса. Использование интерактивных тестов позволяет преподавателям выявить студентов, дающих неверные ответы на тестовые вопросы и в режиме реального времени предоставить для них необходимый контент для изучения и лучшего усвоения учебного материала.

Благодаря достижениям в области информационных технологий, большие данные теперь можно накапливать, анализировать, управлять и в сфере электронного образования. В статье приводится обзор существующих технологий хранения и обработки данных, используемых ведущими странами мира в электронном образовании.

В статье показано, что большие данные в образовании позволяют преподавателям получить разнообразную информацию об уровне подготовки студентов, усвоении учебной информации, выполненных контрольных заданиях и лабораторных работах.

Другой важной проблемой электронного образования являются вопросы выявления новых, порою скрытых, взаимосвязей в больших данных, новых знаний (data mining) [7]. В статье предлагается использовать методы data mining для управления электронным образованием учебными заведениями Азербайджана, для улучшения образовательного процесса и повышения эффективности его управления.

## 1. Data Mining в электронном образовании для анализа больших данных

Технологии Data Mining начали развиваться еще с середины прошлого столетия и лишь в начале этого столетия эти технологии стали использоваться и в области образования. Одним из основателей использования этих технологий в сфере образования является Райян Бэкер (Ryan Baker) — преподаватель Колумбийского университета [8]. Эти технологии получили название EDM (Educational Data Mining). В этот период, в связи с увеличением применения информационных технологий в сфере образования, резко возрастает количество обрабатываемой информации, совершенствуются алгоритмы обработки этой информации.

В основу технологии EDM положена концепция шаблонов (patterns) поведения и личностных качеств обучаемых [9]. Использование этих технологий в сфере образования позволяют узнать какие предметы вызывают у обучаемых большие затруднения, с какими тестами они лучше справляются, какую форму занятий предпочитают, какие темы у них вызывают наибольший интерес и как можно оптимально построить учебную программу, чтобы студент приобрел именно те компетенции, какие ему будут необходимы в сфере его будущей профессиональной деятельности.

Примером использования EDM может быть следующая задача: Имеются ли примеры (шаблоны) полученных оценок выпускниками, которые затем смогли бы в течение короткого времени после окончания вуза найти работу, удовлетворяющую их требованиям.

Основными задачами использования Data Mining в образовании являются:

1. *Классификация* — отнесение объектов (наблюдений, событий) к одному из заранее

известных классов. Для классификации в Data Mining используется множество различных моделей: нейронные сети, деревья решений, метод  $k$ -ближайших соседей [10].

Математически задачу классификации можно записать следующим образом. Имеется некоторое множество  $\{X\}$  описаний (характеристик) объекта, и множество классов  $\{Y\}$ . Существует некоторая целевая функция " $f$ ", преобразующая " $X$ " в " $Y$ " на обучающей выборке  $\{X_m\}$ .

$X_m = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ , где  $x_1, x_2, \dots, x_m$  — векторы признаков объектов, а  $y_1, y_2, \dots, y_m$  — наименование классов, к которым принадлежат соответствующие объекты выборки. Требуется построить алгоритм, преобразующий множество  $X$  в  $Y$ , способный отнести новый (не относящийся к выборке) произвольный объект  $x^0 \in X$  к одному из классов  $Y$ .

Например, степень похожести объектов, а следовательно и вероятность их принадлежности к одному классу, может быть определена на основе расстояния между их точками в пространстве признаков (метод  $k$ -ближайших соседей). Чем меньше расстояние между векторами признаков, тем более похожи друг на друга соответствующие объекты.

Другой метод, используемый для задач классификации — метод деревьев решений (англ.: decision tree), предназначенный для осуществления разбиения исходных данных на группы до тех пор, пока не будут получены множества, состоящие из однородных (схожих) данных. Графически её можно представить в виде древовидной (иерархической) структуры, в узлах (англ. nodes) которых принимаются решения и происходит ветвление (англ.: branching) — деление на ветви (англ.: branches), в зависимости от сделанного выбора. В узлах происходит ветвление процесса, т.е. деление его на так назы-

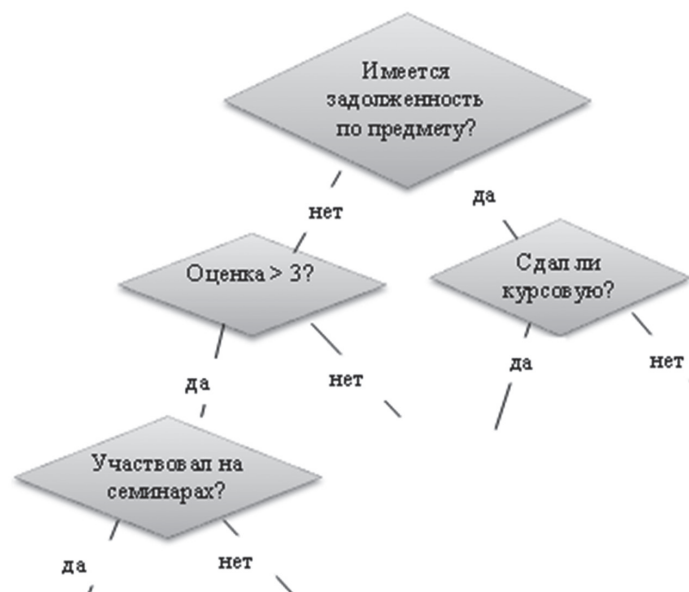


Рис. 1. Пример бинарного дерева решений для задачи классификации

ваемые ветви, а конечные (или, что то же самое, терминальные) узлы называют листьями (англ.: leafs, leaf nodes), в которых принимается конечный результат (решение). Данные конечного узла принадлежат одному классу. На рис. 1 приводится иллюстрация примера использования дерева решений для задач классификации в сфере образования.

Примерами использования задач классификации в электронном образовании являются:

- классификация электронных образовательных ресурсов (по функциональному признаку, определяющему значение и место образовательного ресурса в учебном процессе, по организации текста ресурса, по характеру представляемой информации, по форме изложения, по целевому назначению и т.д.);

- классификация тестовых заданий (по разным уровням сложности, с учётом индивидуального темпа работы, с учетом индивидуальных возможностей обучающегося). При решении этой задачи появляется возможность регулировать количество предъявляемых заданий в зависимости от уровня развития обучающегося.

2. Регрессия, в том числе задачи прогнозирования. приме-

нение регрессионных методов позволит смоделировать влияние, которое оказывают изменение одного из параметров на другой, установить зависимость выходных параметров (целевых функций) от входных переменных (факторов). Благодаря этому можно выяснить, реально ли достичь желаемого результата, если изменить значения выбранного параметра. Решение задач регрессии позволит предсказать результаты итоговых экзаменов, уровень компетенций выпускника, востребованность на рынке труда и уровень их заработной платы после трудоустройства. На основе чего можно выявить степень влияния на процесс образования таких факторов, как, потребность в специалистах; ресурсы вуза (в том числе и финансирование), степень внедрения информационно-телекоммуникационных технологий в образовательный процесс, уровень кадрового обеспечения вуза, заработная плата преподавателей и т.п.

3. Кластеризация — разделение всего множества объектов (наблюдений, событий) которые наиболее близки друг к другу по ряду признаков или свойств, на кластеры. Внутри каждого кластера должны



оказаться похожие объекты, а в разных кластерах — объекты должны отличаться. В результате применения кластерного анализа необходимо решить следующие задачи:

1. Выбрать объекты для кластеризации.
2. Определить множество признаков, по которым будут оцениваться объекты в выборке.
3. Применить один из методов кластерного анализа для создания групп сходных объектов (кластеров).
4. Визуализировать результаты анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Задачу кластеризации в сфере образования можно использовать, например, для обнаружения наличия студентов со сходными психологическими, физиологическими, поведенческими и интеллектуальными характеристиками. Можно выявить как эти поведенческие паттерны (шаблоны) влияют на успешность в различных видах деятельности, какие методы обучения эффективны по отношению к учащимся с различными стереотипами мышления и психики. На основе данных кластеризации можно будет разрабатывать индивидуализированные учебные программы для отдельных групп обучающихся, с учетом длительности обучения, траектории изложения материала, степени сложности заданий и другими характеристиками изучаемой дисциплины.

4. Следующим важным направлением анализа больших данных является *анализ данных социальных сетей*. Дело в том, что большую часть времени учащиеся проводят вне учебных заведений — в социальных сетях, в которых, общаясь со сверстниками, обмениваются информацией, сотрудничают, создают совместный сетевой контент, который в дальней-

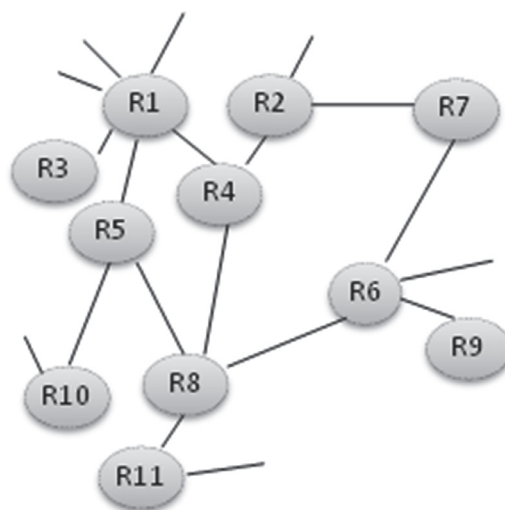


Рис. 2. Фрагмент визуализации взаимодействия студентов в социальной сети

шем может быть использован в образовательном процессе. В связи с тем, что в социальной сети учащиеся находятся в более комфортных условиях, чем в реальности (не находятся под контролем преподавателя), информация, извлеченная из социальных сетей, в отличие от других источников, может быть более объективной. Из социальных сетей можно получить информацию о связях обучающегося со своими сверстниками, о его интересах, перемещениях по сети, активности (частоты вхождения в сеть и времени его нахождения в сети) и др.

В основе анализа социальных сетей лежит математическая теория графов, которая представлена в работе венгерского математика Эрдоса [11]. Математически, сеть — это набор узлов (в нашем примере — это студенты учебных заведений), соединенных линиями, характеризующими отношения между узлами. Каждое отношение соединяет несколько узлов. Первые фундаментальные исследования социальных сетей относятся к 1979 г. и отражены в работах Велмана [12]. Им были разработаны алгоритмы кластерного моделирования и базовые метрики для анализа социальных сетей.

Первым шагом в сетевом анализе является визуализа-

ция данных. Визуализированные графы позволяют выявить узлы, являющиеся ближайшими, найти плотные кластеры активности. Одним из основных параметров графа является «центральность по степени», выражающий отношение количества связей определённого узла к общему количеству других узлов. Если для некоторого узла этот параметр равен 1, это означает, что этот узел связан со всеми остальными узлами сети, если он равен 0, то этот узел является изолированным. Этот показатель показывает степень «знаменитости» узла, показывает, что студент оказывает большое влияние на остальных. На рисунке 2 показан фрагмент визуализированного графа для анализа взаимодействия учащихся в социальной сети. Из рисунка видно, что пользователь сети R1 обладает более высокой степенью центральности, чем пользователь R3 (из R3 отходит меньше линий связи, чем из узла R1)

## 2. Программные приложения для проведения интеллектуального анализа данных в электронном образовании

На сегодняшний день разработано множество программных приложений для интел-

лектуального анализа данных. Так, для задач классификации широко используется алгоритм под названием **C5.0**, представляющий собой стандарт процедуры построения деревьев решений, разработанный программистом J. Ross Quinlan [13]. В интернете имеется бесплатная версия пакета (<http://www.rulequest.com>), реализующего этот алгоритм. С помощью этого пакета можно решить задачи классификации из разных областей, в том числе и из области электронного образования, можно разбить учебную информацию по признакам с большим количеством уровней. Этот пакет может применяться для анализа не только числовых, но и номинальных данных; обеспечивает обработку пропущенных данных. Для построения дерева решений используются только самые важные признаки объектов (выбирает из множества факторов только те, которые сильно влияют на результат классификации). Слабой стороной приложения C5.0 является то, что для задач классификации требуется относительно небольшой объем обучающей выборки, даже небольшие изменения в обучающей выборке сильно влияют на результат. В основном же, этот пакет считается более эффективным, чем другие приложения для классификации.

Для задач классификации широко используется и другой алгоритм под названием CART (classification and regression trees) — для классификации и регрессии с использованием дерева решений [14]. В отличие от C5.0 в алгоритме CART узлы решения имеют две ветки (бинарное представление дерева решений). Результатом классификации может быть, например, ответ на вопрос: «сдаст ли обучаемый успешно экзамен или нет?».

Следующим широко используемым алгоритмом для задач классификации является

метод опорных векторов SVM (Support vector machine). Для задач классификации этот алгоритм использует гиперплоскость для разбиения данных на два класса. Т.е., если у нас имеется набор данных о каждом обучаемом: полученные оценки за семестр, пропущенные занятия, уровень активности на занятиях, степень выполнения домашних заданий и т.д., то можно эти данные использовать как входные в алгоритме. Каждый из этих параметров является измерением в  $n$ -мерном пространстве. SVM отображает эти параметры в этом пространстве и находит гиперплоскость для разделения этих данных на классы. Этот алгоритм используется программным приложением LIBSVM, который можно открыто скачать из сайта <https://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Для задач кластеризации данных широко используются алгоритмы: *k-means* и *c-means*. Первый из этих алгоритмов используется для задач иерархической кластеризации. При иерархической кластеризации, в самом начале, каждый объект помещают в отдельный кластер, далее, эти кластеры объединяют в более крупные кластеры, строится система вложенных разбиений. Т.е. на выходе мы получаем дерево кластеров, корнем которого является вся выборка, а листьями — наиболее мелкие кластеры. Оптимальность разбиения на кластеры определяется по требованию минимизации среднеквадратической ошибки разбиения (*k-means*).

Второй алгоритм используется для нечеткой кластеризации, когда каждому объекту ставится в соответствие набор вещественных значений, показывающих степень отнесения объекта к кластеру с некоторой вероятностью.

Следующим алгоритмом, используемой для целей кластеризации данных является *ЕМ-алгоритм* (*Expectation-*

*Maximization*), являющийся итерационным. На первом шаге итерации (*Expectation*) вычисляется вероятность принадлежности каждой точки данных к кластеру, на втором шаге — (*Maximization*) обновляются параметры модели в соответствии с кластерным распределением, проведенным на предыдущем шаге. На основе ЕМ алгоритма можно сделать предположения о содержании кластера, и о том, в какой кластер следует отнести новые данные. Программную реализацию алгоритма можно найти на сайте <http://www.mathsisfun.com/data/index.html>.

Для анализа социальных сетей (Social Network Analysis, SNA) используется программное приложение *MySpace* [15]. Анализ социальных сетей направлен на изучение отношений между людьми, рассматривает социальное взаимодействие в терминах теории сетей. Эти термины включают в себя узлы (отдельные индивиды внутри сети) и связи или ссылки (представляют отношения между индивидами). *MySpace* может быть использован для интерпретации и анализа структуры и отношений студентов при решении совместных заданий или при взаимодействиях различными средствами коммуникации. При помощи лог-файлов можно отслеживать информацию об обучаемых в социальных сетях (Twitter, Facebook и др.) — измерить количество сообщений в чатах и ссылок на разные разделы изучаемой темы, контакты между участниками и количество вопросов, заданных педагогам.

### 3. Облачные технологии для хранения больших данных электронного образования

Для хранения больших данных требуется большой объем дисковой памяти. Для решения этой проблемы используется кластерная тех-

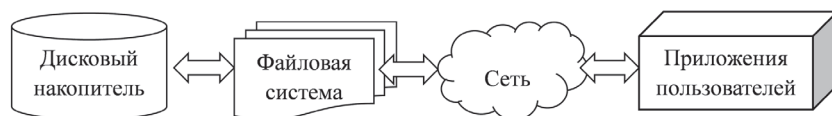


Рис. 3. Типичная архитектура NAS

нология NAS (Network Area Storage), предусматривающая подключение накопителей информации непосредственно к локальной или распределенной компьютерной сети, использующей протокол TCP/IP. Такая сеть позволяет пользователям хранить файлы на NAS серверах и иметь к ним общий доступ с помощью браузера или его сетевого адреса. Операционная система Кластерная инфраструктура NAS состоит из нескольких накопителей (хранилищ) информации, соединенных между собой, позволяет пользователям совместно использовать имеющуюся там информацию и осуществлять поиск. На рисунке 3 представлена архитектура NAS.

По прогнозам специалистов объем рынка NAS-технологий до 2017 года достигнет \$7 млрд. Больше других эти технологии используют США и страны западной Европы, однако в ближайшие годы и в азиатско-тихоокеанском регионе тоже ожидаются темпы роста на уровне 14,1% [16].

Учитывая большой объем информации, которая ежедневно создаются во всем мире, не удивительно, что организации ищут более эффективные и дешевые устройства для хранения данных, какими сейчас являются сетевые хранилища данных, в которую при необходимости можно добавлять новые диски. Однако, уже теперь ясно, что с наступлением эры облачных технологий, анализа больших данных и «Интернета вещей», доступа к данным в режиме реального времени, требует нового подхода как к оптимизации данных внутри храни-

лищ данных, так и доступа к этим данным.

Современные облачные технологии для поддержки требований «Больших данных» по хранению и программному обеспечению предлагают для пользователей оптимизацию хранения, обеспечение безопасности, гибкие методы доставки и масштабируемую инфраструктуру.

В облаках могут находиться не просто огромные объемы необработанных данных, но и данные в их изначальном формате. Новые технологии позволяют их обработать тогда, когда это потребуется. Например, Hadoop, созданный с помощью языка Java, позволяет аналитикам хранить огромные массивы данных, размещая их на большом количестве недорогих серверов, а затем, с помощью MapReduce на виртуальной машине Java (JVM), координировать, объединять и обрабатывать данные [17].

В связи с вышеизложенным, следует отметить совместный проект Мадридского научно-исследовательского института IMDEA Networks, Политехнического университета Мадрида и Университета короля Хуана Карлоса проекта Cloud4BigData [18]. Проект направлен на объединение и интеграцию дифференцированных и специализированных технологий в единую, унифицированную платформу, учитывает требования облачных технологий больших данных, «интернета вещей» и «смарт» — технологий и позволяет в электронном образовании достигнуть: гибкости; масштабируемости; доступности; качества обслу-

живания простоты использования; безопасности и конфиденциальности.

Другими проектами технологии Hadoop являются приложения: **Spark Apache** и **Apache Storm** [19]. Эти приложения позволяют легко и надежно в параллельном режиме обрабатывать неограниченные потоки данных в реальном времени, постепенно увеличивать масштаб и хранить инфраструктуру без потери данных. Spark — проект лаборатории университета Беркли (2009 г. США), служит для распределенной обработки неструктурированных и слабоструктурированных данных. Состоит из ядра и ряда приложений для обработки запросов (Spark SQL), настройки для обработки потоковых распределенных данных (Spark Streaming), набора библиотек машинного обучения (Spark MLlib), программного обеспечения для распределенной обработки графов (GraphX).

Помимо Apache Spark и Storm, для Java-разработчиков представляет интерес еще один проект — DeepLearning4J [20], предназначенный для создания библиотеки машинного обучения для Java и Scala, интегрированную с Hadoop и Spark, может импортировать и данные из сети и создавать многослойную нейронную сеть, использовать для решения задач алгоритмы обучения с учителем и без учителя. Этот проект можно использовать для решения следующих задач электронного обучения: распознавания лиц или изображений; голосовой поиск; распознавание речи (преобразование ее в текст); регрессионный анализ данных и др.

Все эти перечисленные задачи требуют для разрешения обработку больших объемов данных, использующих современные технологии и программные средства, такие, как Map Reduce, NoSQL, Hadoop и другие.



#### 4. Заключение

Концепция больших данных и методы их анализа не так давно стали использоваться для управления предприятиями производственной сферы. Эти технологии могут быть применены и к электронному образованию, включая вопросы принятия решений обработки данных, финансовое планирование и мониторинг успеваемости учащихся. Большие данные позволяют сохранять опыт обучения, дают картину обу-

чения каждого обучающегося. Анализируя эту информацию, при помощи Data Mining, специалисты электронного курса могут разрабатывать индивидуальные траектории обучения и адаптировать процесс обучения для удовлетворения потребностей каждого обучаемого. Анализ данных позволит улучшить модель обучаемого, позволит исследователям получить подробную информацию о характеристиках обучаемого или его состояниях, таких, как знания, мотивация, изучить какие фак-

торы влияют на повышение усвояемости учебного материала, а что мешает этому процессу. Data Mining в электронном образовании позволит педагогам своевременно получать информацию об обучающихся и оперативно реагировать на любые изменения процесса обучения, своевременно вносить изменения в учебный контент.

Мы ожидаем, что рассмотренные в этой статье технологии будут развивать отрасли электронного обучения и в Азербайджане.

#### Литература

1. URL: [www.sloan.com](http://www.sloan.com)
2. Белая книга «Электронное обучение в России» / Под ред. Н.В. Тихомировой, М.: Минобрнауки, 2014. с. 137.
3. Pat Nakamoto. BIG DATA: The revolution that is transforming our work, market and world. Data Analysis / Kindle Edition, 2013. P. 212.
4. Franks B., Taming the big data tidal wave. John Wiley & Sons, Inc. 2010. P. 341.
5. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. М.: Издательство «Манн, Иванов и Фербер», 2014. С. 208.
6. Бадарч Дендев. Информационные и коммуникационные технологии в образовании. М.: ИИТО ЮНЕСКО, 2013. С. 320.
7. Ian H. Witten, Eibe Frank. Data Mining: Practical machine learning tools and techniques. "Morgan Kaufmann", 2015. P. 229.
8. Baker, R. S. Educational data mining: An advance for intelligent systems in education. IEEE Intelligent Systems, 2014. 29 (3). P. 78–82.
9. Bishop Ch. Pattern Recognition and Machine Learning. Series: Information Science and Statistics. 2006. T. XX. 740 p.
10. Дюк В.А., Самойленко А.П. Data Mining: учебный курс. СПб.: Питер, 2001. С. 312.
11. Erdos P., Renyi A. On the evolution of random graphs // Publication of Mathematics Institute Hungary Academy of the Science. 1960. V. 5. P. 17–61.
12. Wellman, B. The community question. American Journal of Sociology, 1979, 84, p. 1201–1207.
13. J. Ross Quinlan: Decision Trees and Instance-Based Classifiers. The Computer Science and Engineering Handbook, 1997. P. 521–535.
14. Зайцева Т. Пусная О. Вероятностные деревья решений: Программная реализация в

#### References

1. URL: [www.sloan.com](http://www.sloan.com)
2. Belaya kniga «Elektronnoe obuchenie v Ros-sii». Ed. N.V. Tikhomirovoy, M.: Minobrnauki, 2014. P.137 (In Russ.)
3. Pat Nakamoto. BIG DATA: The revolution that is transforming our work, market and world. Data Analysis / Kindle Edition, 2013. P. 212.
4. Franks B., Taming the big data tidal wave. John Wiley & Sons, Inc. 2010. P. 341.
5. Mayer-Shenberger V., Kuk'er K. Bol'shie dannye. Revolyutsiya, kotoraya izmenit to, kak my zhivem, rabotaem i myslim. Moscow: Izdatel'stvo «Mann, Ivanov i Ferber», 2014. P. 208.
6. Badarch Dendev. Informatsionnye i kommunikatsionnye tekhnologii v obrazovanii. Moscow: IITO YuNESKO, 2013. P. 320.
7. Ian H. Witten, Eibe Frank. Data Mining: Practical machine learning tools and techniques. "Morgan Kaufmann", 2015. P. 229.
8. Baker, R. S. Educational data mining: An advance for intelligent systems in education. IEEE Intelligent Systems, 2014. 29 (3). P. 78–82.
9. Bishop Ch. Pattern Recognition and Machine Learning. Series: Information Science and Statistics. 2006. T. XX. 740 p.
10. Dyuk V.A., Samoylenko A.P. Data Mining: uchebnyy kurs. Saint Petersburg: Piter, 2001. P. 312. (In Russ.)
11. Erdos P., Renyi A. On the evolution of random graphs. Publication of Mathematics Institute Hungary Academy of the Science. 1960. V. 5. P. 17–61.
12. Wellman, B. The community question. American Journal of Sociology, 1979, 84, p. 1201–1207.
13. J. Ross Quinlan: Decision Trees and Instance-Based Classifiers. The Computer Science and Engineering Handbook, 1997. P. 521–535.
14. Zaytseva T. Pusnaya O. Veroyatnostnye derev'ya resheniy: Programmaya realizatsiya v

решении задач классификации и прогнозирования. Lambert Academic Publisher, 2014. С. 96.

15. URL: <https://myspace.com>

16. Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning. 4(5/6). P. 304–317.

17. Asha T, Shravanthi U.M, Nagashree N, Monika M, Building Machine Learning Algorithms on Hadoop for Bigdata // International Journal of Engineering and Technology. 2013. Vol. 3. No. 2. P. 143–147.

18. Qi Zhang, Lu Cheng, and Raouf Boutaba. Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications. 1(1): 2010. P. 7–18

19. URL: [www.storm.apache.org](http://www.storm.apache.org), URL: [www.spark.apache.org](http://www.spark.apache.org).

20. J. Schmidhuber. Deep learning in neural networks: An overview. Neural Networks, 61: 2015. P. 85–117.

reshenii zadach klassifikatsii i prognozirovaniya. Lambert Academic Publisher, 2014. P. 96 (In Russ.)

15. URL: <https://myspace.com>

16. Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning. 4(5/6). P. 304–317.

17. Asha T, Shravanthi U.M, Nagashree N, Monika M, Building Machine Learning Algorithms on Hadoop for Bigdata. International Journal of Engineering and Technology. 2013. Vol. 3. No. 2. P. 143–147.

18. Qi Zhang, Lu Cheng, and Raouf Boutaba. Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications. 1(1): 2010. P. 7–18

19. URL: [www.storm.apache.org](http://www.storm.apache.org), URL: [www.spark.apache.org](http://www.spark.apache.org).

20. J. Schmidhuber. Deep learning in neural networks: An overview. Neural Networks, 61: 2015. P. 85–117.

#### Сведения об авторах

**Гюляра Абас гызы Мамедова**

Старший научный сотрудник

Институт Информационных Технологий  
Национальной Академии Наук Азербайджана,  
Баку, Азербайджанская Республика

Эл. почта: [gyula.ikt@gmail.com](mailto:gyula.ikt@gmail.com)

Тел.: (994 12) 5397213

**Лала Адхам гызы Зейналова**

Старший научный сотрудник

Институт Информационных Технологий  
Национальной Академии Наук Азербайджана,  
Баку, Азербайджанская Республика

Эл. почта: [lalamaillala@bk.ru](mailto:lalamaillala@bk.ru)

**Рена Тофик гызы Меликова**

Старший научный сотрудник

Институт Информационных Технологий  
Национальной Академии Наук Азербайджана,  
Баку, Азербайджанская Республика

Эл. почта: [rena22@rambler.ru](mailto:rena22@rambler.ru)

Тел.: (994 12) 5397213

#### Information about the authors

**Gyulara A. Mamedova**

Senior researcher

Azerbaijan National Academy of Sciences, Institute of  
Information Technology,  
Baku, Republic of Azerbaijan

E-mail: [gyula.ikt@gmail.com](mailto:gyula.ikt@gmail.com)

Tel.: (994 12) 5397213

**Lala A. Zeynalova**

Senior researcher

Azerbaijan National Academy of Sciences, Institute of  
Information Technology, Baku, Republic of Azerbaijan

E-mail: [lalamaillala@bk.ru](mailto:lalamaillala@bk.ru)

Tel.: (994 12) 5397213

**Rena T. Melikova**

Senior researcher

Azerbaijan National Academy of Sciences, Institute of  
Information Technology,  
Baku, Republic of Azerbaijan

E-mail: [rena22@rambler.ru](mailto:rena22@rambler.ru)

Tel.: (994 12) 5397213