ISSN 2708-2032
e-ISSN 2708-2040

INTERNATIONAL UNIVERSITY

# INTERNATIONAL JOURNAL OF INFORMATION & COMMUNICATION TECHNOLOGIES

Volume 2, Issue 2
June, 2021

ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН
MINISTRY OF EDUCATION AND SCIENCE OF THE REPUBLIC OF KAZAKHSTAN


# INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

# МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

# ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

Том 2, Выпуск 2
Июнь, 2021

Издание зарегистрировано Министерством информации и общественного развития Республики Казахстан. Свидетельство о постановке на учет № KZ82VPY00020475 от 20.02.2020 г.

Журнал зарегистрирован в Международном центре по регистрации сериальных изданий ISSN (ЮНЕСКО, г. Париж, Франция)

Выходит 4 раза в год.

УЧРЕДИТЕЛЬ:
АО «Международный университет информационных технологий»

ISSN 2708-2032 (print)
ISSN 2708-2040 (online)

## СОДЕРЖАНИЕ

### РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ И ИНЖЕНЕРИЯ ЗНАНИЙ

***Бактаев А.Б., Мукажанов Н.К.***
Алгоритм решения задачи по исправлению опечаток в тексте, применяемый в поисковых системах с поддержкой казахского языка ..... 9
***Еркетаев Н.М., Мукажанов Н.К.***
Эффективное хранение неструктурированных данных ..... 19
***Сагадиев Р.Т., Шайкемелев Г.Т.***
Представление логической витрины данных в экосистеме Hadoop ..... 28
***Бейсенбек Е.Б., Дузбаев Н.Т.***
Современные способы взлома и защиты ПО ..... 33
***Найзабаева Л.К., Алашыбаев Б.А.***
Рекомендательная система для онлайн-магазинов с использованием машинного обучения ..... 38
***Мейрамбайулы Н., Дузбаев Н.Т.***
Мониторинг стационарных источников выбросов загрязняющих веществ г. Алматы ..... 47

### ИНФОКОММУНИКАЦИОННЫЕ СЕТИ И КИБЕРБЕЗОПАСНОСТЬ

***Айтмагамбетов А.З., Кулакаева А.Е., Койшыбай С.С., Жолшибек И.Ж.***
Исследование возможностей применения низкоорбитальных спутников для радиомониторинга в республике Казахстан ..... 54
***Кемельбеков Б.Ж., Полуанов М.***
Анализ метода бриллюэновской рефлектометрии в волоконно-оптических линиях связи ... 62
***Турбекова К.Ж.***
Анализ применения БПЛА в сетях связи при чрезвычайных ситуациях ..... 68

### ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

***Азанов Н.П., Хабиров Р.Р., Әміров У.Е.***
Конкурентная разведка и принятие решений с помощью машинного обучения для обеспечения промышленной безопасности ..... 75
***Джаныбекова С.Т., Толганбаева Г.А., Сарсембаев А.А.***
Распознавание говорящего с помощью глубокого обучения ..... 85
***Салерова Д.К., Сарсембаев А.А.***
Обзорная статья распознавания номерных знаков с использованием оптического распознавания символов ..... 93
***Салерова Д.К., Сарсембаев А.А.***
Исследование существующих методов классификации изображений ..... 100
***Оразалин А., Мурсалиев Д.Е., Сергазина А.С.***
Актуальные сверточные архитектуры нейронной сети для диагностики медицинских изображений ..... 115
***Әлімхан А.М.***
Прогнозирование результатов игры в баскетбол с использованием алгоритмов глубокого обучения ..... 112

***Адырбек Ж.А., Сатыбалдиева Р.Ж.***
Анализ процессов планирования и решения проблем в логистике с помощью интеллектуальной системы ..... 120
***Нургалиев М.К., Алимжанова Л.М.***
Геймификация в образовании ..... 128
***Базарбеков И.М., Шарипов Б.Ж.***
Разработка бизнес-процесса для получения онлайн услуг в организации образования ..... 133
***Жунусов Д.О., Алиаскаров С.Ж.***
Метод классификации текстов на основе алгоритмов машинного обучения ..... 140

**ЦИФРОВЫЕ ТЕХНОЛОГИИ В ЭКОНОМИКЕ И МЕНЕДЖМЕНТЕ**

***Алимжанова Л.М., Панарина А.В.***
Внедрение сервисной системы IT-аутсорсинга ..... 146
***Жұмабай Р.Ж., Алимжанова Л.М.***
Управление процессами работы с поставщиками на основе ERP-стандартов — подход BPM ..... 153
***Бердыкулова Г.М., Төлепбергенова Д.А.***
Менеджмент университета: практика МУИТ ..... 159
***Омарова А.Ш., Алимжанова Л.М., Таштамышева А.Э.***
Исследование и разработка методов перехода традиционного маркетинга в цифровой формат ..... 166

## CONTENTS

### SOFTWARE DEVELOPMENT AND KNOWLEDGE ENGINEERING

***Baktayev A.B., Mukazhanov N.K.***
Algorithm for solving the problem of correcting typos with search engines supporting the Kazakh language ..... 9
***Yerketayev N.M., Mukazhanov N.K.***
Efficient storage of unstructured data ..... 19
***Sagadiyev R.T., Shaikemelev G.T.***
Representing a logical data mart in the Hadoop ecosystem ..... 28
***Beisenbek Y.B., Duzbaev N.T.***
Modern methods of hacking and protection software ..... 33
***Naizabayeva L., Alashybayev B.A.***
A recommendation system for online stores using machine learning ..... 38
***Meirambaiuly N., Duzbaev N.T.***
Monitoring of stationary sources of pollutant emissions in Almaty ..... 47

### INFORMATION AND COMMUNICATION NETWORKS AND CYBERSECURITY

***Aitmagambetov A.Z., Kulakayeva A.E., Koishybai S.S., Zholshibek I.Z.***
Study of the possibilities of using low-orbit satellites for radio monitoring in the Republic of Kazakhstan ..... 54
***Kemelbekov B.J., Poluanov M.***
Analysis of the brillouin reflectometry method in fiber-optic communication lines ..... 62
***Turbekova K.Zh.***
Analysis of the use of UAVs in emergency communication networks ..... 68

### SMART SYSTEMS

***Azanov N.P., Khabirov R.R., Amirov U.E.***
Competitive intelligence and decision-making algorithm using machine learning for industrial security ..... 75
***Janybekova S.T., Tolganbayeva G.A., Sarsembayev A.A.***
Speaker recognition using deep learning ..... 85
***Salerova D.K., Sarsembayev A.A.***
Review of license plate recognition using optical character recognition ..... 93
***Salerova D.K., Sarsembayev A.A.***
Research on the existing image classification methods ..... 100
***Orazalin A., Mursaliyev D.E., Sergazina A.S.***
Current convolutional neural network architectures for diagnosing medical images..... 105
***Alimkhan A.M.***
Predicting basketball results using deep learning algorithms ..... 112
***Adyrbek Zh.A., Satybaldiyeva R.Zh.***
Analysis of the planning and problem-solving processes in logistics using an intelligent system ..... 120
***Nurgaliyev M.K., Alimzhanova L.M.***
Gamification in education ..... 128

***Bazarbekov I.M., Sharipov B.Zh.***
Development of a business process for obtaining online services in the organization of education ..... 133
***Zhunissov D.O., Aliaskarov S.Zh.***
Method for text classification based on machine learning algorithms ..... 140

**DIGITAL TECHNOLOGIES IN ECONOMICS AND MANAGEMENT**

***Alimzhanova L.M., Panarina A.V.***
Implementation of an IT outsourcing service system ..... 146
***Zhumabay R.Zh., Alimzhanova L.M.***
Supplier process management based on ERP standards: the BPM approach ..... 153
***Berdykulova G.M., Tolepbergenova D.A.***
University management: case study of IITU ..... 159
***Omarova A.Sh., Alimzhanova L.M., Tashtamysheva A.E.***
Research and development of methods for the transition of traditional marketing to digital format ..... 166

## МАЗМҰНЫ

### БАҒДАРЛАМАЛЫҚ ҚАМТАМАНЫ ӘЗІРЛЕУ ЖӘНЕ БІЛІМ ИНЖЕНЕРИЯСЫ


***Бактаев А.Б., Мукажанов Н.К.***
Қазақ тілін қолдайтын іздеу жүйелерінде қолданылатын мәтіндегі жаңылыстарды түзету бойынша есептерді шешу алгоритмі ..... 9
***Еркетаев Н.М., Мукажанов Н.К.***
Құрылымсыз деректерді тиімді сақтау ..... 19
***Сагадиев Р.Т., Шайкемелев Г.Т.***
Hadoop экожүйесінде логикалық деректер кесіндісін ұсыну ..... 28
***Бейсенбек Е.Б., Дузбаев Н.Т.***
Бағдарламалық жасақтаманы бұзудың және қорғаудың заманауи әдістері ..... 33
***Найзабаева Л., Алашыбаев Б.А.***
Машиналық оқытуды қолдану арқылы интернет-дүкендерге арналған ұсыныс жүйесі ..... 38
***Мейрамбайұлы Н., Дузбаев Н.Т.***
Алматы қаласы бойынша ластаушы заттар шығарындыларының стационарлық дереккөздеріне мониторинг жүргізу ..... 47


### АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ЖЕЛІЛЕР ЖӘНЕ КИБЕРҚАУІПСІЗДІК


***Айтмагамбетов А.З., Кулакаева А.Е., Койшыбай С.С., Жолшибек И.Ж.***
Қазақстан Республикасында радиомониторинг үшін төмен орбиталық спутниктерді қолдану мүмкіндіктерін зерттеу ..... 54
***Кемельбеков Б.Ж., Полуанов М.***
Талшықты-оптикалық байланыс желілеріндегі бриллюэн рефлектометрия әдісін талдау ... 62
***Турбекова К.Ж.***
Төтенше жағдайлар кезінде байланыс желілерінде ПҰА-ның қолданылуын талдау ..... 68


### ИНТЕЛЛЕКТУАЛДЫ ЖҮЙЕЛЕР


***Азанов Н.П., Хабиров Р.Р., Әміров У.Е.***
Өнеркәсіптік қауіпсіздікті қамтамасыз ету үшін машиналық оқытуды қолдана отырып, бәсекеге қабілеттілікті барлау және шешім қабылдау ..... 75
***Джаныбекова С.Т., Толғанбаева Г.А., Сарсембаев А.А.***
Терең оқыту арқылы сөйлеушіні тану ..... 85
***Салерова Д.К., Сарсембаев А.А.***
Таңбаларды оптикалық тануды пайдалану арқылы нөмірлер белгілерін тануға шолу мақаласы ..... 93
***Салерова Д.К., Сарсембаев А.А.***
Қолданыстағы бейнелерді жіктеу әдістерін зерттеу ..... 100
***Оразалин А., Мурсалиев Д.Е., Серғазина А.С.***
Медициналық кейіндік диагностикаға арналған конволюциялық жүйкелік желі архитектурасы ..... 105
***Әлімхан А.М.***
Терең оқыту алгоритмдерін қолдана отырып, баскетбол нәтижелерін болжау ..... 112

***Адырбек Ж.А., Сатыбалдиева Р.Ж.***
Логистикадағы жоспарлау процестерін талдау және логистикадағы интеллектуалды жүйені қолдану арқылы мәселелерді шешу ..... 120
***Нұрғалиев М.Қ., Алимжанова Л.М.***
Білім беру саласындағы геймификация ..... 128
***Базарбеков И.М., Шарипов Б.Ж.***
Білім беру ұйымында онлайн қызмет көрсету үшін бизнес-процессін дамыту ..... 133
***Жунусов Д.О., Алиаскаров С.Ж.***
Машинналық оқыту алгоритмдері негізінде мәтіндер классификациясының әдісі ..... 140

## ЭКОНОМИКА ЖӘНЕ БАСҚАРУДАҒЫ САНДЫҚ ТЕХНОЛОГИЯЛАР

***Алимжанова Л.М., Панарина А.В.***
IT-аутсорсингтің сервистік жүйесін енгізу ..... 146
***Жұмабай Р.Ж., Алимжанова Л.М.***
ERP стандарттарына негізделген жеткізушілермен жұмыс процесін басқару - BPM тәсілі ..... 153
***Бердыкулова Г.М., Төлепбергенова Д.А.***
Университетті басқару: ХАТУ практикасы ..... 159
***Омарова А.Ш., Алимжанова Л.М., Таштамышева А.Э.***
Дәстүрлі маркетингті цифрлық форматқа ауыстыру әдістерін зерттеу және әзірлеу ..... 166

# РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ И ИНЖЕНЕРИЯ ЗНАНИЙ

УДК 004.021

**Бактаев А.Б.[1], Мукажанов Н.К.[2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан

## АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ ПО ИСПРАВЛЕНИЮ ОПЕЧАТОК В ТЕКСТЕ, ПРИМЕНЯЕМЫЙ В ПОИСКОВЫХ СИСТЕМАХ С ПОДДЕРЖКОЙ КАЗАХСКОГО ЯЗЫКА

**Аннотация.** В статье описывается подход решения и применения алгоритма исправления опечаток или ошибок в словах с поддержкой казахского языка, которые допустили люди при вводе текста. Алгоритм разрабатывается с помощью расстояния Дамерау – Левенштейна и редакционного предписания для поисковой системы. Для точности и релевантности поиска, требуется точные ключевые запросы. В спешке люди могут забыть переключить раскладку, допустить орфографические или другие виды ошибок при наборе текста, и нельзя допустить чтобы это повлияло на качество поиска информации.

**Ключевые слова**: исправление опечатка в тексте, поисковые запросы, ошибки в запросах, расстояние Дамерау — Левенштейна, редакционное предписание.

### Введение

Количество информационных ресурсов, таких как новостные сайты, блоги, площадки электронной коммерции с онлайн продажами, электронные энциклопедии, непрерывно растет. Существует мнение, что поисковая машина для сайта не является важным элементом, заслуживающим внимания и траты ресурсов. Возможно, это правда в каких-то частных ситуациях. Однако в большинстве случаев поисковая машина играет важную роль в жизненном цикле сайта. Данная работа посвящена одной из главных частей поисковой системы — предварительной обработке текстового запроса для улучшения релевантности поиска. Необходимость предварительной обработки запроса перед выполнением поиска обусловлена технологической особенностью поискового процесса: в большинстве случаев поисковые машины используют полнотекстовый поиск, поэтому чем точнее выполнен запрос, тем точнее будет результат.

Принцип предварительной обработки текста можно реализовать с помощью алгоритма редакционного предписания. Также можно применить алгоритм расстояния Левенштейна для определения разности между двумя последовательностями символов. Требуется база данных существующих слов, чтобы вычислительная система понимала, какие слова существуют в естественном языке. Естественный язык включает в себя более тысячи различных языков со своими правилами, синтаксисом и т.д. За основу возьмем три языка: казахский, русский, английский. Люди при наборе текста допускают чаще всего ошибки следующего характера:

- самый распространенный вариант — проявление человеческого фактора — опечатка, которую возможно допустить где угодно. Например, «сегдня сонлечный день». В данном предложении ошибки в словах сегодня и солнечный;
- слова из русского языка, набранные латинскими буквами, или наоборот. Забыли сменить раскладку клавиатуры. Например, «Privet» или «Ghbdtn» вместо «Привет»;
- казахский текст, написанный в русской раскладке. Например, «ынгайлы» вместо «ыңғайлы», или «коркем ан» вместо «көркем ән».

**Основная концепция решения задачи по исправлению опечаток**

Чтобы получить нужную информацию, найти ответ, необходимо правильно сформулировать поисковой запрос. Результат поиска напрямую зависит от введенного для поиска текста. Текст является набором слов, а слова — последовательностью символов, которая имеет определенное значение. Поскольку мыслительная деятельность протекает быстрее физических действий, при наборе текста пользователи могут допустить ошибку или опечатку. Наша цель — по возможности найти и исправить эти ошибки, чтобы результат поиска всегда был более релевантным.

Предположим, что имеется база данных с миллионами слов и одно слово на вход для того, чтобы найти корректный исходный вариант. Поиск совпадения путем сопоставления длины нужного слова с каждым словом в базе данных, высчитывая сходство между $s_1$ и $s_2$, — слишком дорогая и долгая операция. Извлечь данные из памяти быстрее, чем выполнять трудоемкие вычисления при каждом запросе. Все дело в реализации данного функционала. Вместо того чтобы перебирать процент совместимости с каждым словом в базе данных, можно предварительно подготовить набор возможных вариантов с помощью алгоритма редакционного предписания и поискать совпадения в проиндексированной базе данных. Если результаты от предыдущих действий не удовлетворительны, можно расширить диапазон с помощью смены раскладки между кириллицей и латиницей, воспользоваться снова алгоритмом редакционного предписания и попытаться найти совпадения в базе данных уже с другой раскладкой.

Перебирая возможные варианты с помощью редакционного предписания, получаем в результате кандидатов для исправления. Для определения одного варианта, нуждающегося в корректировке, можно использовать множество вариантов разных метрик и подходов. В данной работе сравним теорию вероятности с алгоритмом расстояния Левенштейна.

**Алгоритм**

**Дистанция редактирования** между двумя строками (редакционное расстояние) — это наименьшее количество операций вставки, удаления и замены одного символа на другой для превращения одной строки в иную [1].

**Расстояние Дамерау-Левенштейна** — это расширение еще одной операцией алгоритма расстояния Левенштейна. Операция называется транспозицией, перестановка рядом расположенных символов местами. Алгоритм хорошо работает с однобайтными кодировками, но на некоторых языках допустимо описание текстов только с использованием кодировок с переменной длиной символов, например Юникод [2]. Он выявил, что при наборе текста 4/5 ошибок являются транспозициями [3].

**Редакционное предписание** — это порядок действий для получения из одной строки второй наикратчайшим образом. Чаще всего действия обозначаются так, как указано в таблице «Обозначения операций редакционного предписания» (таблица 1) [1]. Для операций вставки (insert) используются буквы из алфавита языка, на котором требуется найти слово.

Таыблица 1 – Обозначения операций редакционного предписания

| Операция | Полное название на английском | Полное название на русском |
|---|---|---|
| **D** | Delete | удалить |
| **I** | Insert | вставить |
| **R** | Replace | заменить |
| **T** | Transpose | транспозиция |
| **M** | Match | совпадение |

К примеру, для строк «СӘЛЕМ» и «САЕМ» можно построить преобразование как показано в таблице «Пример работы редакционного предписания» (таблица 2) [3]. Первые

символы совпадают, обозначаем буквой M, что означает Match (совпадение) согласно таблице 1. Следующие операции — замена, вставка, затем снова два совпадения.

Таблица 2 – Пример работы редакционного предписания

| **Операция** | **M** | **R** | **I** | **M** | **M** |
|---|---|---|---|---|---|
| ***Корректное слово*** | С | Ә | Л | Е | М |
| ***Слово с ошибкой*** | С | А | … | Е | М |

Также из слова СӘЛЕМ можно построить разные варианты с ошибками, используя редакционное предписание (рисунок 1).

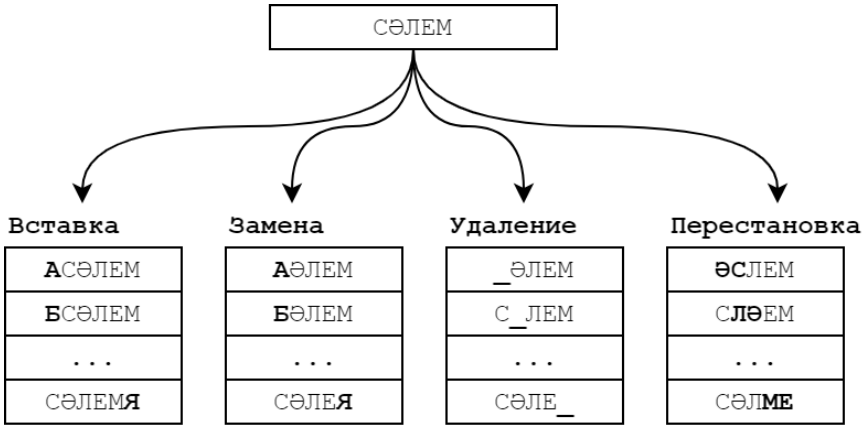



*Рисунок 1 - Демонстрация работы редакционного предписания*

Можем считать, что все $\omega$ неотрицательны: если две последовательные операции можно заменить одной, это не ухудшает общую цену (к примеру, заменить символ $x$ на $y$, а потом с $y$ на $z$ не лучше, чем сразу $x$ на $z$) [1].

Очевидно, справедливы следующие утверждения:

$d(s_1,s_2) \geq ||s_1| - |s_2||$

$d(s_1,s_2) \leq \omega(\varepsilon,b) * |s_1| + \omega(a,\varepsilon) * |s_2|$

$d(s_1,s_2) = 0 \leftrightarrow s_1 = s_2$

где $d(s_1,s_2)$ — расстояние с подстановкой между $s_1$ и $s_2$, $|s|$ — длина строки $s$ [1].

Основаным минусом является то, что алгоритм требует $O(M * N)$ операций и точно такую же память. Чем длиннее строка, тем дольше операция будет выполняться. Это значит, что потребуется примерно 40 гигабайт памяти для сравнения файлов длиной в $10^5$ строк [2]. Учитывая, что областью применения является поисковая система, долгое выполнение не приемлемо. По этой причине мы выставим ограничения по длине каждого слова. Предположительно она будет равна длине самого длинного слова в базе данных.

**Определение кандидата с помощью теорий вероятности** — по каждому найденному слову-кандидату ищем частоту повторения в базе данных и делим его на количество всех слов в базе данных, и возвращаем кандидата, где получили максимальное значение.

**Смена раскладки текста** — это преобразование текста с кириллицы на латиницу или наоборот в соответствии с раскладкой на клавиатуре (рисунок 2).

Допустимо наличие нескольких раскладок для одного письменного языка. Например, имеются раскладки ЙЦУКЕН и фонетическая (ЯВЕРТЫ) для русского языка. QWERTY, Дворака и Colemak для английского языка [4].

Клавиатурная раскладка казахского языка разрабатывалась на основе русской раскладки и закреплена стандартом РСТ КазССР 903-90. Раскладка казахского, русского и английского (рисунок 2) [4].


*Рисунок 2 - Клавиатурная раскладка казахского, русского и английского языка*

Для смены раскладки создадим коллекцию, которая представляет из себя пару ключ-значения (таблица 3).

Таблица 3 – Коллекция ключ-значения для смены раскладки

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Latin** | q | w | e | r | t | y | u | i | o | p | [ | ] |
| **Cyrillic** | й | ц | у | к | е | н | г | ш | щ | з | х | ъ |
| **Latin** | Q | W | E | R | T | Y | U | I | O | P | { | } |
| **Cyrillic** | Й | Ц | У | К | Е | Н | Г | Ш | Щ | З | Х | Ъ |
| | | | | | | | | | | | | |
| **Latin** | a | s | d | f | g | h | j | k | l | ; | ' | |
| **Cyrillic** | ф | ы | в | а | п | р | о | л | д | ж | э | |
| **Latin** | A | S | D | F | G | H | J | K | L | : | " | |
| **Cyrillic** | Ф | Ы | В | А | П | Р | О | Л | Д | Ж | Э | |
| | | | | | | | | | | | | |
| **Latin** | z | x | c | v | b | n | m | , | . | | | |
| **Cyrillic** | я | ч | с | м | и | т | ь | б | ю | | | |
| **Latin** | Z | X | C | V | B | N | M | < | > | | | |
| **Cyrillic** | Я | Ч | С | М | И | Т | Ь | Б | Ю | | | |
| | | | | | | | | | | | | |
| **Latin** | @ | # | $ | % | * | ( | ) | _ | + | | | |
| **Cyrillic** | Ә | І | Ң | Ғ | Ү | Ұ | Қ | Ө | h | | | |
| **Latin** | 2 | 3 | 4 | 5 | 8 | 9 | 0 | - | = | | | |
| **Cyrillic** | ә | і | ң | ғ | ү | ұ | қ | ө | h | | | |

**Предварительная обработка текста (корректировка текста)** с поддержкой казахского языка решается с помощью комплекса алгоритмов и подходов для решения задачи (рисунок 3). Данный подход учитывает смену раскладки клавитатуры, различные варианты опечаток и т.д.

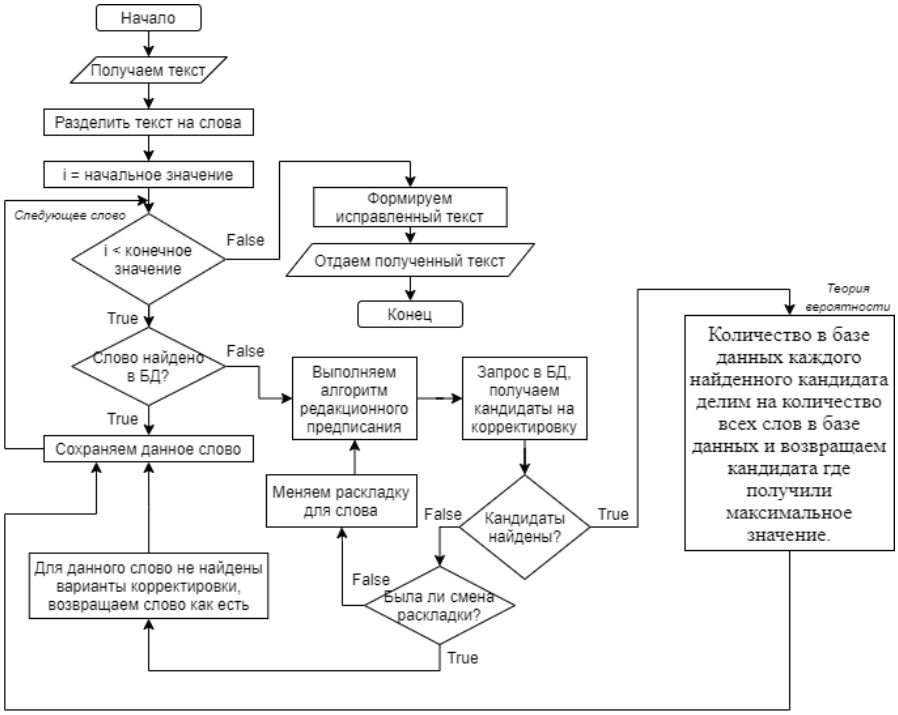



*Рисунок 3 - Блок-схема алгоритма исправления корректировки текста*

**Шаги алгоритма корректировки текста** (рисунок 3):

1) входной текст разделяем на отдельные слова;
2) начинаем перебирать каждое слово;
3) сначала пробуем найти соответствие в базе данных без каких-либо изменений, чтобы не совершать лишних операций в случае, если слово изначально верное;
4) если находим слово в базе данных, добавляем в итоговую коллекцию;
5) если совпадения в базе данных по этому слову не найдено, выполняется алгоритм редакционного предписания, на выходе получаем набор возможных вариантов, по которым ищем кандидатов в базе данных слов;
6) если кандидаты не найдены, проверяем, была ли ранее смена раскладки, например, с английского на русский, казахский или наоборот. Если была, меняем раскладку, тем самым расширив диапазон поиска кандидатов на корректировку. Далее возвращаемся на предыдущий шаг, снова выполняем редакционное предписание и пытаемся получить кандидатов на корректировку;
7) если после применения алгоритма редакционного предписания количество кандидатов равно нулю, исходное слово возвращаем в качестве корректного слова, так как в базе данных отсутствует данное слово и кандидаты на его корректировку;
8) если кандидаты найдены и при этом количество кандидатов больше одного, возникает новая задача: определить, какой вариант из подобранных кандидатов подходит на корректировку. У данной задачи есть различные способы решения. В данной статье рассмотрим теорию вероятности, алгоритм расстояния Дамерау-Левенштейна;
9) по полученным результатам слова из коллекций конкатенируем и возвращаем в качестве предварительно обработанного текста для поисковой системы.

Сделать выбор из полученных кандидатов на корректировку — задача непростая. Данную проблему можно решить с помощью различных метрик и даже с помощью теории вероятности. Вопрос в том, какую цель преследовать: скорость или точность?

**Анализ методов**

Если сравнить теорему Байеса и расстояние Левенштейна, то отдать предпочтение какому-либо одному методу сложно. Попробуем сравнить точность и время выполнения (таблица 4).

Таблица 4 – Сравнение расстояния Левенштейна и теории вероятности в практическом применении на языке программирования python

| **Метрика Левенштейна** | **Теория вероятности** |
|---|---|
| **Text:** кайрлы<br>**Elapsed time:** 0.2219 seconds<br>**Результат:** қайырлы | **Text:** кайрлы<br>**Elapsed time:** 0.2248 seconds<br>**Результат:** жайлы |
| **Text:** интеркт<br>**Elapsed time:** 0.2866 seconds<br>**Результат:** интернет | **Text:** интеркт<br>**Elapsed time:** 0.2897 seconds<br>**Результат:** интернет |
| **Text:** кайырлы кун алматы каласы<br>**Elapsed time:** 0.0004 seconds<br>**Результат:** қайырлы<br>**Elapsed time:** 0.0002 seconds<br>**Результат:** қайырлы күн<br>**Elapsed time:** 0.0002 seconds<br>**Результат:** қайырлы күн алматы<br>**Elapsed time:** 0.0003 seconds<br>**Результат:** қайырлы күн алматы баласы | **Text:** кайырлы кун алматы каласы<br>**Elapsed time:** 0.0003 seconds<br>**Результат:** қайырлы<br>**Elapsed time:** 0.0002 seconds<br>**Результат:** қайырлы күн<br>**Elapsed time:** 0.0002 seconds<br>**Результат:** қайырлы күн алматы<br>**Elapsed time:** 0.0003 seconds<br>**Результат:** қайырлы күн алматы баласы |

**Тестирования метода подбора кандидатов**

Множество вариантов зависит от базы данных слов. Чем больше слов, тем больше вариантов, и тем сложнее определить точно, что имел в виду пользователь.

Рассмотрим проблемы обоих методов. Ниже представлен список из введенных для корректировки слов и их варианты. Например, для слова *карай* нашлось 5 кандидатов на казахском языке. По метрике расстояния Левенштейна, у всех вариантов метрика вычислила одинаковое расстояние «1 — на один символ каждый из них отличается от исходного». Выбор в данном случае будет очевидно случайным. Теорией вероятности в данном случае будет учитываться, насколько часто используется в базе данных это слово. Если частота выше, чем у остальных кандидатов, алгоритм подберет то слово, которое используется чаще всего.

**Например варианты:**

**Карай** - {'арай', 'қадай', 'жасай', 'асай', 'тарай', 'атай', 'жарай', 'маңай', 'санай', 'малай', 'шавай', 'тақай', 'калай', 'ламай', 'шалай', 'сағай', 'сарай', 'адай', 'апай', 'жанай', 'алай', 'надай', 'абай', 'кабак', 'талай', 'ағай', 'қарай'}

**Сут** - {'суы', 'сәт', 'сот', 'суыт', 'сат', 'су', 'сүт', 'вут'}

**Калам** - {'қалам', 'далам', 'балам', 'салам', 'шалам', 'алам'}

**Кате** - {'кете', 'катер', 'күте', 'қате', 'кафе'}

**Каласы** - {'қаласы', 'баласы', 'наласы'}

**Карай** - {'сарай', 'арай', 'тарай', 'қарай', 'жарай'}

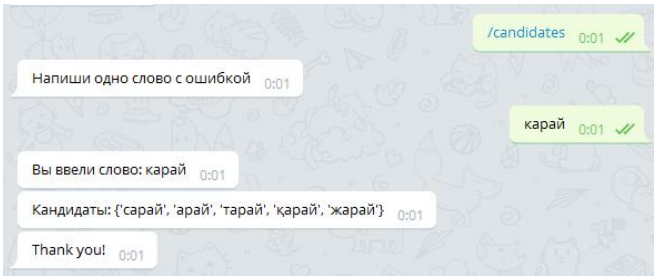



*Рисунок 4 - Бот для корректировки текста (подбор кандидатов)*

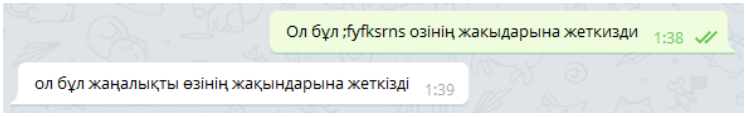



*Рисунок 5 - Бот для корректировки текста (исправление предложения)*

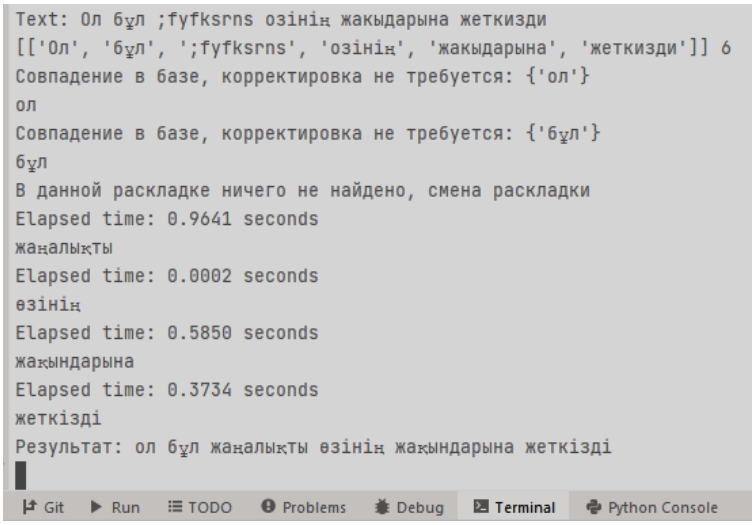



*Рисунок 6 - Консоль и время выполнения корректировки каждого слова*

**Заключение**

**Проведенная работа**

В данной статье были описаны основные подходы для решения задачи корректировки текста на казахском языке. В частности, расширение казахскими буквами редакционного предписания и его применение, расширение базы казахскими словами, анализ и сравнение метрики Левенштейна и теории вероятности, смена раскладки текста, демонстрация примера на блок-схеме и результатов экспериментов с казахскими словами и предложениями, написанными на языке python.

**Результаты**

После сравнения теории вероятности и расстояния Левенштейна можно понять, что расстояние Левенштейна дает более релевантный результат, чем теория вероятности. Однако, какую бы метрику не использовали для определения, разницей между словами *интернет* и *интернат* будет один символ, что делает определение достаточно сложным. Для улучшения точности дополнительно нужно учитывать контекст запроса. Соответственно, в базе данных нужно построить контекстные зависимости для слов.

**Планы на будущее**

Опечатки можно исправлять методами нечеткого поиска и линейным поиском [5]. На сегодняшний день не существует алгоритмов, которые смогли бы с точностью в 100 %

определить, какое слово имел в виду человек, но к этому нужно стремиться. Возможно, применение машинного обучения в сфере лингвистики поможет сделать определение того, что имел в виду человек в своем тексте, более точным.

Чаще всего применяются различные виды структур нейронных сетей для повышения точности распознавания, которые требуют сложного длительного обучения и значительной обучающей выборки [6]. Распознование некоторых текстов может быть ограничено по сложности, по алфавиту, по расположению, однако качество корректировки зависит от качества обучения нейронной сети [6]. Но есть и обратная сторона этой проблемы — скорость и цена процесса распознавания ошибок в тексте. Однако разработчики Яндекса показали на практике, применив CatBoost, разработанный их компанией, что использование машинного обучения увеличивает точность корректировки [7].

Для улучшения данного функционала как части поисковой системы можно применить OLAP (online analytical processing) по данным поисковых запросов людей, совместно с машинным обучением, а также нормализацию текста с помощью NLP.

В то же время база слов должна постоянно автоматизированно снабжаться новыми словами, так как появляются новые слова и сленг. Кроме того, нужно учитывать переход на латинский алфавит казахского языка, перевод базы данных казахских слов на латиницу для работы алгоритма, описанного в этой статье.

## СПИСОК ЛИТЕРАТУРЫ


1. Бадалов М. И., Мирзамов А. М. Редакционное расстояние с подстановкой //toshkent shahridagi turin politexnika universiteti. – 2017. – С. 304.
2. Сидоркина И. Г., Килеев В. В. Кодировка символов переменной длины в алгоритме Дамерау–Левенштейна //Вестник Чувашского университета. – 2013. – №. 3.
3. Расстояние Левенштейна. [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Расстояние_Левенштейна. (дата обращения: 23.04.2021)
4. Раскладка клавиатуры. [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Раскладка_клавиатуры. (дата обращения: 23.04.2021)
5. Бондаренко А. О. Библиотека для исправления опечаток с учетом контекста: дис. – Сибирский федеральный университет, 2020.
6. Сапаров А. Ю., Бельтюков А. П., Маслов С. Г. Уточнение результатов распознавания математических формул с использованием расстояния Левенштейна //Вестник Удмуртского университета. Математика. Механика. Компьютерные науки. – 2020. – Т. 30. – №. 3. – С. 513-529.
7. Салахутдинова К. И., Лебедев И. С., Кривцова И. Е. Алгоритм градиентного бустинга деревьев решений в задаче идентификации программного обеспечения //Научно-технический вестник информационных технологий, механики и оптики. – 2018. – Т. 18. – №. 6.


## REFERENCES


1. Badalov M. I., Mirzamov A. M. Redakcionnoe rasstoyanie s podstanovkoi //toshkent shahridagi turin politexnika universiteti. – 2017. – S. 304.
2. Sidorkina I. G., Kileev V. V. Kodirovka simvolov peremennoi dliny v algoritme Damerau–Levenshteina //Vestnik Chuvashskogo universiteta. – 2013. – №. 3.
3. Rasstoyanie Levenshtejna. [Electronic resource] URL: https://ru.wikipedia.org/wiki/Расстояние_Левенштейна. (date of the application: 23.04.2021)
4. Raskladka klaviatury. [Electronic resource] URL: https://ru.wikipedia.org/wiki/Раскладка_клавиатуры. (date of the application: 23.04.2021)
5. Bondarenko A. O. Biblioteka dlya ispravleniya opechatok s uchetom konteksta : dis. – Sibirskij federalnyi universitet, 2020.

6. Saparov A. U., Beltyukov A. P., Maslov S. G. Utochnenie rezultatov raspoznavaniya matematicheskih formul s ispolzovaniem rasstoyaniya Levenshteina //Vestnik Udmurtskogo universiteta. Matematika. Mekhanika. Komputernye nauki. – 2020. – T. 30. – №. 3. – S. 513-529.
7. Salahutdinova K. I., Lebedev I. S., Krivsova I. E. Algoritm gradientnogo bustinga dereviev reshenii v zadache identifikacii programmnogo obespecheniya //Nauchno-tekhnicheskii vestnik informacionnyh tekhnologii, mekhaniki i optiki. – 2018. – T. 18. – №. 6.

**Бактаев А.Б., Мукажанов Н.К.**

**Қазақ тілін қолдайтын іздеу жүйелерінде қолданылатын мәтіндегі жаңылыстарды түзету бойынша есептерді шешу алгоритмі**

**Аңдатпа.** Қазақ тілін қолдайтын, мақалада мәтінді енгізу кезінде сөздердегі қателермен жаңылыстарды түзету тәсілдері және алгоритмі сипатталған. Алгоритм Дамерау-Левенштейн қашықтығымен және іздеу жүйесіне арналған редакциялық нұсқаулық арқылы жасалады. Іздеудің дәлдігі мен өзектілігі үшін нақты негізгі сұраулар қажет. Асығыста адамдар пернетақта орналасуын ауыстыруды, теру кезінде орфографиялық немесе басқа да қателіктердің пайда болуы мүмкін, бірақ бұл ақпаратты іздеу сапасына әсер етуге жол бермеу керек.

**Түйінді сөздер**: Мәтіндегі қатені түзету, іздеу сұраулары, сұраулардағы қателер, Дамерау – Левенштейн қашықтығы, редакциялық нұсқаулық.

**Baktayev A.B., Mukazhanov N.K.**

**Algorithm for solving the problem of correcting typos with search engines supporting the Kazakh language**

**Abstract.** This article describes an approach to solving and applying an algorithm for correcting typos and spelling mistakes which people make when entering a text in the Kazakh language.

The algorithm is developed using the Damerau-Levenshtein distance and editorial prescription for search system (engine). An exact keyword (query) is required for more accurate and relevant search. In a hurry, people tend to forget to switch the keyboard layout, make spelling or other types of mistakes when entering a text, and this should not be allowed to affect the quality of information retrieval.

**Keywords:** correction of typos in text, search queries, query errors, Damerau-Levenshtein distance, editorial prescription.

**Авторлар туралы мәлімет:**

**Бактаев Айдос Бакдаулетович**, бакалавр, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті.

**Мукажанов Нуржан Какенович**, PhD, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының ассистент-профессоры, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Бактаев Айдос Бакдаулетович**, бакалавр, магистрант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**Мукажанов Нуржан Какенович**, PhD, ассистент-профессор кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**About the authors:**

**Aidos B. Baktayev**, master student, Department of Computer Engineering and Information Security, International Information Technology University.

**Nurzhan K. Mukazhanov**, PhD, Assistant-Professor, Department of Computer Engineering and Information Security, International Information Technology University.

УДК 004.651:623

**Еркетаев Н.М. [1], Мукажанов Н.К. [2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан

# ЭФФЕКТИВНОЕ ХРАНЕНИЕ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ

**Аннотация.** В этой статье представлены результаты изучения того, как могут быть использованы различные типы информации для хранения неструктурированных данных в базе данных и классической файловой системе. Также объясняются современные методы хранения данных, предыстория проблемы и ключевые элементы, важные для этого исследования: различные методы хранения неструктурированных данных (преимущества и недостатки), используемые типы данных и экспериментальная среда. Наша основная цель — найти эффективный метод хранения неструктурированных данных.

**Ключевые слова**: база данных, неструктурированные данные, файловый поток, эффективность, производительность.

## Введение

Хорошо известно, что одним из результатов быстрого роста Интернета стало значительное увеличение объема информации, генерируемой и распространяемой организациями практически во всех отраслях и секторах. Проблема не только в генерации данных, но и в их хранении и доступе к ним. Менее известна степень потребности в больших объемах дорогостоящих ресурсов, как человеческих, так и технических, обусловленной происходящим информационным взрывом. Эти потребности, в свою очередь, создали столь же большую, но в значительной степени неудовлетворенную потребность в инструментах, которые можно использовать для управления тем, что мы называем неструктурированными данными. Следует признать, что термин неструктурированные данные может означать разные вещи в разных контекстах. Например, в контексте систем реляционных баз данных это относится к данным, которые нельзя хранить в строках и столбцах. Вместо этого такие данные должны храниться в BLOB (большом двоичном объекте), универсальном типе данных, доступном в большинстве программных систем управления реляционными базами данных (RDBMS). Здесь под неструктурированными данными понимаются файлы электронной почты, текстовые документы, презентации, файлы изображений и видеофайлы.

С другой стороны, стремительный рост объемов данных обусловлен достижениями и серьезными преобразованиями в технологии магнитной записи. Стоимость хранения резко упала с более чем 4 долларов за мегабайт в 1990 году до менее чем 0,001 доллара за мегабайт сегодня.

Согласно исследованию, проведенному IDC, ведущей фирмой по исследованию и анализу рынка информационных технологий, объем данных, которые будут собираться, храниться и воспроизводиться по всему миру, вырастет со 161 эксабайта в 2006 году до 988 эксабайт в 2010 году (1 эксабайт = 1018 байт). Два ключевых вывода этого исследования:

- Основная часть этих данных будет в виде изображений, снятых большим количеством устройств, таких как цифровые камеры, телефоны с камерами, камеры наблюдения и медицинское оборудование для визуализации. Большая часть этих данных должна храниться и управляться централизованными системами внутри организаций. Исследование показывает, что к 2010 году, хотя предприятия будут создавать, собирать и тиражировать только 30% цифровой вселенной [3], им придется хранить более 85% всех данных и управлять ими.
- Более 95% цифровой вселенной — это неструктурированные данные. Согласно этому исследованию, 80% всех хранимых организациями данных не структурированы.

Ожидается, что эта тенденция роста сохранится и в будущем, поскольку потребует эффективных способов хранения, поиска, структурирования и обеспечения безопасности неструктурированных данных. Об аналогичных тенденциях в отношении важности обработки неструктурированных данных также сообщили другие исследовательские и консультационные фирмы, такие как Gartner Group и Butler Group [4].

Очевидно, что неструктурированные данные — это наша реальность, и поиск эффективного метода хранения данных является ключевым аспектом этого исследования и будущих результатов в этой области. В настоящей статье объясняются современные методы хранения данных.

**Связанная работа**

Существует множество исследований, основанных на тестировании систем баз данных. Одна статья, представляющая собой хорошую отправную точку для нашей работы, особенно интересна: «Набор тестов для обработки неструктурированных данных» [2].

Основной целью авторов было собрать набор рабочих нагрузок, выявить и изучить широкий спектр приложений для обработки неструктурированных данных, выявить их ключевые характеристики обработки и доступа к вводу-выводу, а также составить рабочие нагрузки, воплощающие эти характеристики. Поэтому, чтобы спроектировать системы хранения для этого важного развивающегося класса приложений, они создают набор тестов, который может фиксировать их характеристики обработки и ввода-вывода.

Пакет Benchmark состоит из четырех рабочих нагрузок:

- Обнаружение края;
- Поиск близости;
- Сканирование данных;
- Слияние данных.

Был сделан вывод, что приложения, использующие неструктурированные данные для бизнес-процессов, очень интенсивны по вводу-выводу и предъявляют высокие требования к системе хранения [2].

Но предварительные исследования оставляют возможность продолжить подобные эксперименты с неструктурированными данными.

**История исследований**

В следующем разделе будут объяснены предыстория и ключевые элементы, важные для этого исследования: различные методы хранения неструктурированных данных (преимущества и недостатки), используемые типы данных и экспериментальная среда. Наша основная цель — найти эффективный метод хранения неструктурированных данных. В ходе этого процесса мы проведем обширный сравнительный анализ на основе четко определенных параметров.

Исследование основано на следующих методах хранения неструктурированных данных:

- Неструктурированные данные в среде реляционных баз данных;
- Неструктурированные данные вне реляционной среды.

Мы начнем с объяснения этих методов, каждого преимущества и побочных эффектов.

*A. Неструктурированные данные в реляционной среде*

Спор обычно возникает относительно темы реляционных баз данных и данных больших двоичных объектов (BLOB): интегрировать большие двоичные объекты в базу данных или хранить их в файловой системе? Каждый метод имеет свои преимущества и недостатки.

Хранение неструктурированных данных, таких как изображения, аудиофайлы и исполняемые файлы в базе данных с типичными текстовыми и числовыми данными, позволяет вам хранить вместе всю связанную информацию для данного объекта базы данных

(рис. 1). И этот подход позволяет легко искать и извлекать данные BLOB; вы просто запрашиваете соответствующую текстовую информацию. Однако хранение неструктурированных данных может значительно увеличить размер ваших баз данных.

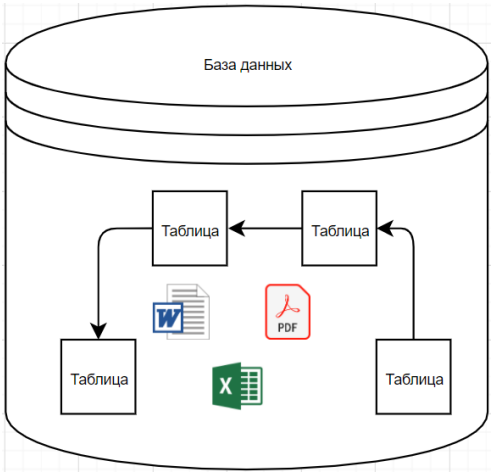



*Рисунок 1 - Среда базы данных с неструктурированными данными внутри*

Таблица-1. Характеристики базы данных

| | Характеристики базы данных | | | | |
|---|---|---|---|---|---|
| | Назначение базы данных | Размер с неструктурированными данными | Количество пользователей | Количество BLOB-записей | Время резервного копирования |
| 1. | LMS/CMS | ≈ 12 ГБ | ≈ 5000 студентов | ≈ 7000 файлов размером от 100 Кб до 15 Мб | 25 мин. |

**Основные преимущества использования этого метода:**

- Когда данные хранятся в базе данных, резервная копия является согласованной. Нет необходимости в отдельной политике резервного копирования;
- Когда данные хранятся в базе данных, это часть транзакции. Так, например, откат включает все традиционные операции с базой данных вместе с операциями с двоичными данными. Обычно это делает клиентское решение более надежным и с меньшим количеством кода.

**Основные недостатки использования этого метода:**

- Хранение неструктурированных данных позволяет значительно увеличить размер баз данных;
- Резервное копирование и восстановление может занять много времени;
- Проблемы с производительностью подсистем ввода-вывода;

*Б. Неструктурированные данные вне реляционной среды*

Распространенной альтернативой вышеописанному методу является хранение двоичных файлов вне базы данных с последующим включением в качестве данных в базу данных пути к файлу или URL-адреса объекта.

Этот отдельный метод хранения имеет несколько преимуществ по сравнению с интеграцией данных BLOB в базе данных. Это несколько быстрее, потому что чтение данных из файловой системы требует немного меньше накладных расходов, чем чтение

данных из базы данных. А без больших двоичных объектов базы данных, как правило, меньше.

Однако нам необходимо вручную создать и поддерживать связь между базой данных и файлами внешней файловой системы, которые могут рассинхронизироваться. Кроме того, обычно нам требуется уникальная схема именования или хранения файлов ОС, чтобы четко идентифицировать потенциально сотни или даже тысячи файлов BLOB [1].

Хранение данных BLOB в базе данных устраняет эти проблемы, позволяя хранить данные BLOB вместе с соответствующими реляционными данными (рис. 2).

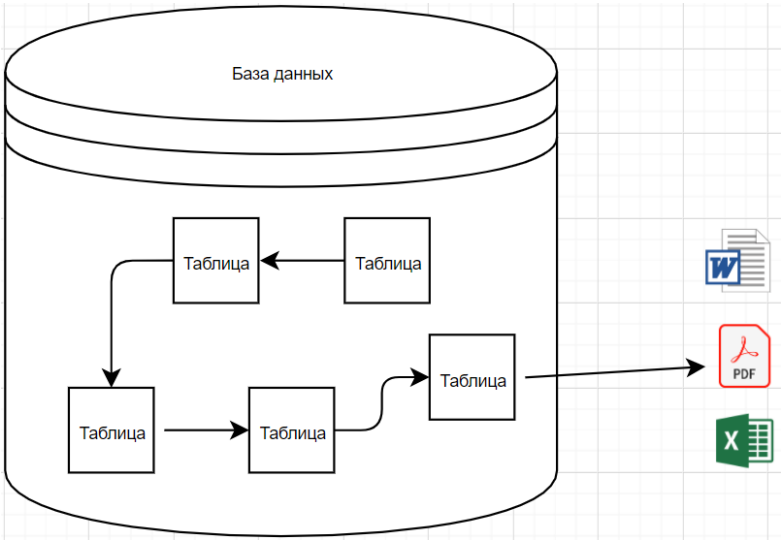



*Рисунок 2 - Среда базы данных без неструктурированных данных внутри*

В таблице 2 у нас такая же среда базы данных, как и в таблице 1. Единственное отличие состоит в том, что мы исключаем пару таблиц, которые содержат неструктурированные данные внутри физического файла базы данных. Разница более чем очевидна: размер базы данных при одинаковом количестве пользователей составляет половину от первоначального. Давайте подробнее рассмотрим преимущества этого метода.

Таблица-2. Характеристики базы данных

| | Характеристики базы данных | | | | |
|---|---|---|---|---|---|
| | Назначение базы данных | Размер с неструктурированными данными | Количество пользователей | Количество BLOB-записей | Время резервного копирования |
| 1. | LMS/CMS | ≈ 6 ГБ | ≈ 5000 студентов | 0 | 11 мин. |

**Основные недостатки использования этого метода:**

- Когда данные хранятся вне базы данных, резервная копия не согласована;
- Неструктурированные данные не являются частью транзакции.

**Основные преимущества использования этого метода:**

- Хранение неструктурированных данных за пределами базы данных может снизить размер баз данных и снизить пропускную способность ввода-вывода;

- Резервное копирование и время восстановления могут занять меньше времени;

*C. Гибридный способ хранения неструктурированных данных*

Проблема с первыми двумя методами заключается в том, что мы не знаем, как фактический размер данных влияет на производительность базы данных. Результаты пока не говорят нам, есть ли разница в хранении BLOBS: 10 КБ, 1 МБ, 100 МБ и т. д.

Основные поставщики баз данных теперь поддерживают гибридный способ хранения неструктурированных данных. В этом случае данные находятся «вне» среды базы данных. Но главное преимущество заключается в том, что BLOB объекты находятся в условиях транзакционной согласованности базы данных (рис. 3).

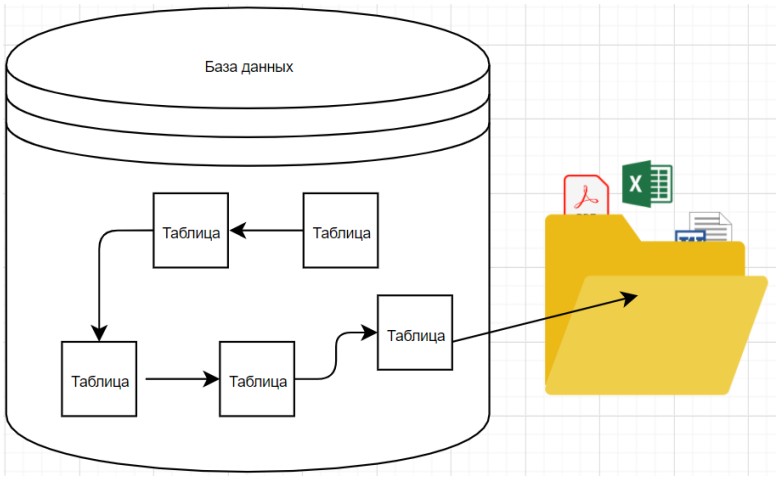



*Рисунок 3 - Гибридный способ хранения неструктурированных данных*

В нашем случае используются те типы данных, которые предлагаются в новой версии SQL Server 2008. Использование этой новой технологии фактически делает возможным гибридный метод хранения неструктурированных данных. Ядро базы данных поддерживает новый тип данных с именем filestream.

Ключевым аспектом исследования является выяснение эффективности этого метода хранения данных.

**Экспериментальная среда**

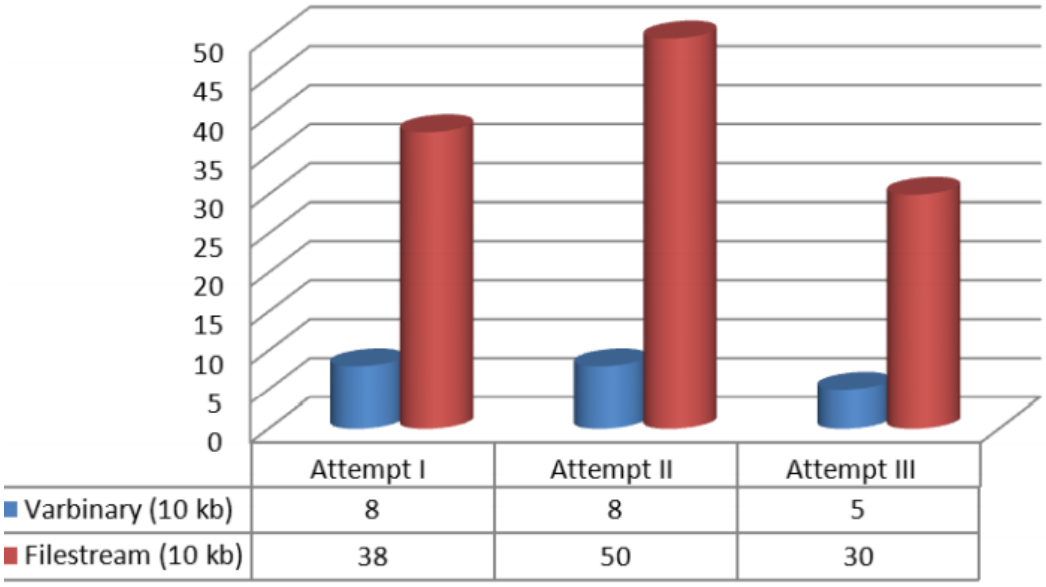
Для этого мы настраиваем тестовую среду со следующими компонентами:

- Процессор: AMD X2 3.0 Ghz 4 ГБ;
- Физическая память;
- Файлы базы данных на C;
- Компонент файлового потока на диске E;
- Диски C: и E: на отдельных физических дисках SATA;
- SQL Server 2008 R2 и клиентское приложение для настраиваемого тестирования находятся на одном компьютере;

На следующих диаграммах показано среднее время загрузки для следующих условий:

- Файл размером 10 КБ, повторяется 3 раза для каждого измерения (рис. 4);
- Файл размером 1 МБ, повторяется 3 раза для каждого измерения (рис. 5);
- Файл размером 10 МБ, повторяется 3 раза для каждого измерения (рис. 6);

**Результаты для файла размером 10 КБ**

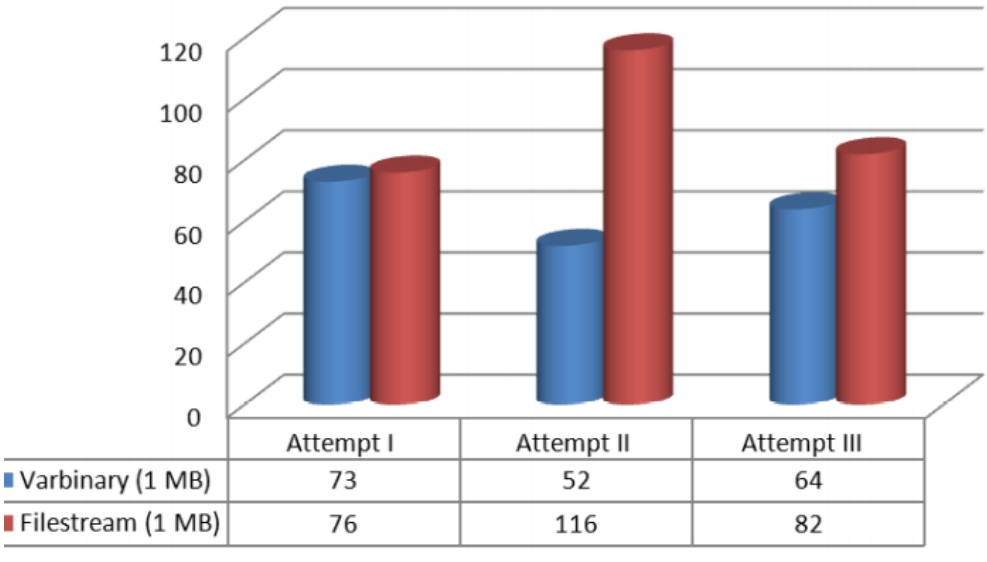



| | Attempt I | Attempt II | Attempt III |
|---|---|---|---|
| Varbinary (10 kb) | 8 | 8 | 5 |
| Filestream (10 kb) | 38 | 50 | 30 |

*Рисунок 4 - Результаты сохранения файла размером 10 КБ внутри базы данных и файловой системы.*

В этом измерении вы можете четко увидеть накладные расходы и влияние на производительность, вызванные использованием файлового потока на “маленькие” файлы. Время при хранении внутри базы данных было каждый раз меньше 10 миллисекунд. С другой стороны, время хранения с использованием файлового потока в 4-5 раз больше.

**Результаты для файла размером 1 МБ**




| | Attempt I | Attempt II | Attempt III |
|---|---|---|---|
| Varbinary (1 MB) | 73 | 52 | 64 |
| Filestream (1 MB) | 76 | 116 | 82 |

*Рисунок 5 - Результаты сохранения файла размером 1 МБ внутри базы данных и файловой системы.*

При объеме данных 1 МБ традиционный двоичный файл и файловый поток действуют одинаково, и разница между ними — максимум на один раз быстрее.

**Результаты для файла размером 10 МБ**

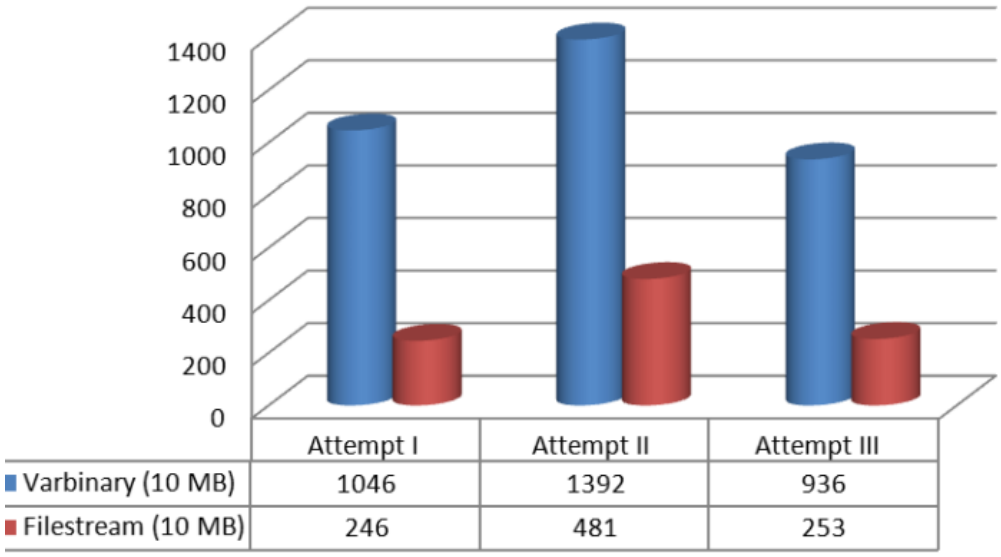



*Рисунок 6 - Результаты сохранения файла размером 10 МБ внутри базы данных и файловой системы.*

Когда используются файлы размером 10 МБ, хранение данных в традиционном файле базы данных происходит намного медленнее. Основываясь на этих измерениях, можно сказать, что более эффективно использовать файловый поток, когда типичный размер файла составляет около 1 МБ или более. Если файлы имеют небольшой размер (явно менее 1 МБ), традиционное хранилище работает лучше. При измерении времени мы обнаружили, что удаление строк на основе файлового потока происходит намного быстрее, чем при хранении внутри таблицы.

**Заключение**

В исследовании представлена гипотеза о неэффективности существующих методов хранения и извлечения неструктурированных данных.

Новый способ использования файловой системы при согласованности базы данных был хорошей возможностью опробовать новую технологию в разных областях хранения данных. Примеры тестов сохранения файлов размерами 10 КБ, 1 МБ и 10 МБ в реальных средах информационных систем могут относиться к фотографиям пользователей, файлам резюме, небольшим офисным документам, видео - и аудиофайлам. Результаты исследования могут использоваться для моделирования системы и четкого определения потребностей в хранении данных. Преимущество заключается в получении максимальной производительности на основе аппаратной и программной инфраструктуры. Также в системах, в которых проблемы с производительностью уже существуют, эта модель может помочь выявить узкие места и найти способ их улучшить.

Результаты исследования позволяют создать модель тестирования системы для хранения неструктурированных данных. Дальнейшие шаги по улучшению — реализация этой модели в реальных средах, где важны неструктурированные данные (платформы электронного обучения, социальные сети, порталы хранения видео и т. д.). После этого настоящее исследование может стать официальной методологией анализа системы с потребностями в неструктурированных данных. Современные тенденции [3] показывают нам, что www становится глобальным хранилищем для всех видов данных, где преобладают неструктурированные данные.

СПИСОК ЛИТЕРАТУРЫ


1. Hitachi Global Storage Technologies - Обзорные схемы технологии HDDT. [Электронный ресурс] URL:http://www.hitachigst.com/hdd/technolo/overview/stor agetechchart.html. (дата обращения: 10.02.2021)
2. Набор тестов для обработки неструктурированных данных Smullen, C.W.; Tarapore, S.R.; Gurumurthi, S.; Архитектура сети хранения данных и Параллельный ввод-вывод, 2007г. SNAPI. Международный семинар 24 сентября. 2007 Страниц: 79 – 83.
3. J. Gantzandetal. Расширение цифровой вселенной - прогноз роста мировой информации до 2010 г., март 2007 г. Белая книга IDC.
4. C.White. Консолидация, доступ и анализ неструктурированных данных, O'Reilly Media, 2005 г. Страницы 118–124.
5. Р. Чемберлен, М. Франклин. Использование реконфигурируемости для текстового поиска. В материалах семинара по высокопроизводительным встроенным вычислениям (HPEC), O'Reilly Media, 2006 г. Страницы 1178–1181.
6. Повышение доступности данных с помощью файловых сетей Geer, B; Компьютер. O'Reilly Media, 2017 г. Страницы 1356–1359.
7. Основы классификации неструктурированных данных Островски, Д.А.; Семантические вычисления, 2009. ICSC '09. Международная конференция IEEE, 14–16 сентября 2009 г. Страницы: 373–377.
8. Анализ неструктурированных данных и оптимизация их хранения. [Электронный ресурс] URL: https://habr.com/ru/company/hpe/blog/265499. (дата обращения: 19.02.2021)
9. Преобразование неструктурированных данных из разрозненных источников в знания Plejic, B.; Vujnovic, B.; Penco, R.; Семинар по приобретению знаний и моделированию, 2008 г. Семинар по KAM, 2008 г. Международный симпозиум IEEE 21–22 декабря 2008 г. Страницы: 924 – 927.
10. G.Fountainand и S.Drager. Высокопроизводительная архитектура слияния в реальном времени. O'Reilly Media, 2002 г. Страницы 1478–1485.


REFERENCES


1. Hitachi Global Storage Technologies – Overview diagrams of HDDT technology. [Electronic resource] URL:http://www.hitachigst.com/hdd/technolo/overview/stor agetechchart.html. (date of the application: 10.02.2021)
2. A Benchmark Suite for Unstructured Data Processing Smullen, C.W.; Tarapore, S.R.; Gurumurthi, S.; Storage Network Architecture and Parallel I/Os, 2007. SNAPI.International Workshop on 24 Sept. 2007 Page(s): 79 – 83
3. J.Gantzandetal. The Expanding Digital Universe – A Forecast of Worldwide Information Growth Through 2010, March 2007. IDC Whitepaper
4. C.White. Consolidating, Accessing, and Analyzing Unstructured Data, O'Reilly Media, 2005 Page(s): 118 – 124
5. R. Chamberlain, M. Franklin and R. Index. Using reconfigurability for text search. In High Performance Embedded Computing (HPEC) Workshop Proceedings. O'Reilly Media, 2006 Page(s): 1178 – 1181
6. Increasing data availability with Geer, D. File networks; Computer. O'Reilly Media, 2017 Page(s): 1356 – 1359
7. Fundamentals of unstructured data classification Ostrowski, D.A.; Semantic Computing, 2009. ICSC '09. IEEE International Conference, September 14-16, 2009, Pages: 373-377
8. Analysis of unstructured data and optimization of their storage. [Electronic resource] URL: https://habr.com/ru/company/hpe/blog/265499. (Date accessed: 19.02.2021)
9. Converting unstructured data from disparate sources to knowledge Plejic, B.; Vujnovich, B.; Penco, R .; Knowledge Acquisition and Modeling Workshop 2008 KAM Workshop 2008 IEEE International Symposium December 21-22, 2008 Pages: 924 - 927

10. J.Fountain and S.Drager. High-performance real-time merge architecture. O'Reilly Media, 2002 Page(s): 1478 – 1485

**Н.М. Еркетаев, Н.К. Мұқажанов**
**Құрылымсыз деректерді тиімді сақтау**

**Аңдатпа.** Бұл мақалада құрылымсыз деректерді мәлімет базасында және классикалық файлдық жүйеде сақтау үшін сан түрлі типтегі деректерді пайдалану бойынша зерттеу нәтижелерін ұсынамыз. Ол сонымен қатар деректерді сақтаудың заманауи әдістерін, фондық тарихты және негізгі элементтерді түсіндіреді. Осы зерттеуге қатысты: құрылымдалмаған мәліметді сақтаудың әртүрлі әдістері (артықшылықтары мен кемшіліктері), қолданылатын мәлімет типтері және эксперименттік орта. Біздің басты мақсат – құрылымдалмаған деректерді сақтаудың тиімді әдісін табу.

**Түйінді сөздер**: мәлімет қоры, құрылымдалмаған мәлімет, файлдар ағыны, тиімділік, өнімділік.

**N.M. Yerketayev, N.K. Mukazhanov**
**Efficient storage of unstructured data**

**Abstract.** In this article we present the research findings on the use of different types of unstructured data stored in a database and the classic file system. It also explains modern data storage methods, background history and key elements. The issues relevant to this research: different methods of storing unstructured data (advantages and disadvantages), the used data types and the experimental environment. Our main goal is to find an efficient method for storing unstructured data.

**Keywords:** database, unstructured data, file stream, efficiency, performance.

**Авторлар туралы мәлімет:**

**Еркетаев Нұрзат Мейірханұлы**, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының магистранті, Халықаралық ақпараттық технологиялар университеті.

**Мұқажанов Нуржан Какенұлы**, PhD, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының ассистент-профессоры, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Еркетаев Нұрзат Мейірханұлы**, магистрант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**Мукажанов Нуржан Какенович**, PhD, ассистент-профессор кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**About the authors:**

**Nurzat M. Yerketayev,** master student, Department of Computer Engineering and Information Security, International Information Technology University.

**Nurzhan K. Mukazhanov**, PhD, Assistant-Professor, Department of Computer Engineering and Information Security, International Information Technology University.

УДК 530.1, 681.3.06

**Sagadiyev R.T.[1], Shaikemelev G.T.[2]**

[1,2] International Information Technology University, Almaty, Kazakhstan

# REPRESENTING A LOGICAL DATA MART IN THE HADOOP ECOSYSTEM

**Abstract.** This article reveals the concept of a logical data mart within the Hadoop Big Data ecosystem, which is used in production for analytical purposes and business decision making. The article provides an in-depth look at the modern approach to processing Big Data, and clearly demonstrates the use of logical data marts with the help of diagrams and images. Upon completion of the work, a conclusion is made about the work performed, including a summary of the main ideas of the article.

**Keywords:** technology, logical data mart, data warehouse, information processing, advantages, Hadoop.

## Introduction

Today, due to the desire to provide the best service on the market, many large enterprises turn to modern technologies, introducing them into their work processes. To achieve this goal, you need to have valuable information that can be extracted from any area of interaction with the client and other parties, and then, based on data analytics, make an important decision for the business. Sources of this information can be Internet resources, corporate information (archives, transactions and databases), as well as readings from reading devices, such as cellular sensors, etc. Since such sources have recently become very popular, whole terabytes of raw information are formed.

Such a large amount of data is referred to among the IT community as Big Data, which comprises a growing bulk of both structured and unstructured data generated every second [1]. To accomplish the task of processing Big Data, traditional data analysis tools will not be enough, therefore, new technologies for storing large amounts of data have been developed, which, in turn, make it possible to create logical data marts that allow enterprises to consolidate all the data they have.

A prime example and a key project for Big Data processing is the Apache Software Foundation project named as the Apache Hadoop ecosystem. The Hadoop project is optimized for running on virtual machines and scales easily, making it suitable for any business handling data volumes in excess of a terabyte. Therefore, it is the best option to develop such a tool as a logical big data mart within the Hadoop ecosystem, which allows you to link a huge amount of information from different sources into a single view without creating a single physical storage.

## Concept of logical data marts

Big Data technologies are intended to offer methods and other means for transmitting massive volumes of data to big IT infrastructures. However, the sheer amount of data available today might not be the only issue. In enterprise and other serious applications, you often work on large amounts of data with a dynamic and fluid structure, as well as disparate collections of data. There are problems the solutions for which are unknown in advance, and the researcher needs software to examine the raw data or the outcomes of equations based on it without a programmer’s assistance. One way to get such a tool is to create a logical data mart.


*Figure 1 - The areas of Big Data application*

To explain the definition of a logical data mart in the context of this article, it is important to first clarify what it is. A logical data mart is a collection of decomposed subsets of data in the form of sliced arrays of thematically narrowly oriented information tailored to the needs of a particular user community (Fig. 1) [2]. The data mart enables the user to get answers to critical market issues and greatly expands the possibilities of learning from the company's data.

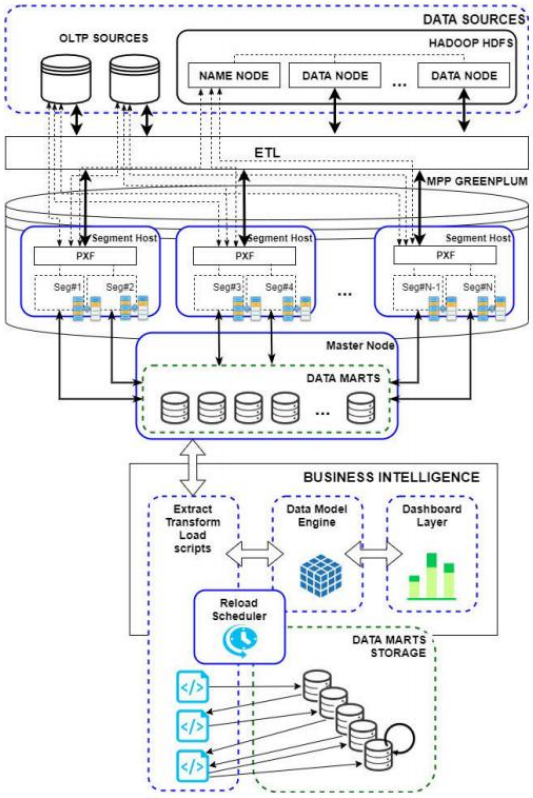



*Figure 2 - Conceptual view of logical data marts' information structure*

In traditional modeling, any object in the physical world can be represented as a single multi-argument table. A data source is a set of minimally redundant, collectively stored interconnected tables.

A data source table may be designated as part of a knowledge resource to which a community of users requires online access. As a result, the data mart has access to each source through the data source table, is "aware" of its structure, and is able to request data from the source structures, consolidate it in a single structure, automatically merge data from various sources and present it to

the user (Fig. 2). The customer, in turn, will send questions to which an exact answer must be supplied, rather than a list of searched terms, to the data mart, which the latter interprets as a logical expression.

The key functions of a logical data mart can be distinguished from the above:

1) the logical data mart should be able to retrieve any data from any source upon request.

2) it should be able to index weakly structured sources, such as file storages, and create a structured semantic description for each document found in them.

3) the user should be able to "ask questions" to the logical data mart, which should have the most accurate, not approximate answers to such questions.

4) the user should not have to be concerned with the data's exact location or configuration.

5) the logical data mart is not required to store all data in one storage.

The logical data mart must "talk" with the user in his language to "understand" the context of the questions, so a formalized explanation of the subject area is needed. If a company deals with pipelines, metering systems, and readings, for example, these concepts should be used in the technical model that the system uses.

**Logical data mart in Hadoop**

The processing of big amounts of digital information is one of the problems that stands in the way of even the largest IT-companies. It is very hard to perform analytics via big data marts using the standard methodologies and software tools due to a large size and complexity of computational tasks. It is very difficult for a single machine to efficiently deal with very large datasets. Therefore, the past decade has seen a rapid development of the big data technologies in many areas of our life. Distributed computing technology has been the subject of development for many modern IT-companies. Using scalable cluster computing systems instead of single powerful machines is the thing which has been invented with the aim to deal with big data.

In addition to the challenges of data processing and methods of extracting useful information, the storage of large amounts of data marts is a fundamental issue. Notwithstanding a large amount of data storage architectures that have been developed over the years, in today's world where petabytes of data are generated every day, there is still a growing need for many organizations to rethink their approaches to data storage in favor of new, highly scalable systems.

Rather than providing a project-oriented or subject-area-based data center, more organizations are adopting the idea of a data center or data pool, where data is collected and stored based on source networks. The advantage of this strategy is that it allows you to remodel and build data marts whenever you choose, depending on your needs. The Hadoop Big Data framework can handle conventional batch systems as well as provide a near-real time decision support (2 seconds to 2 minutes after data delivery) or near-real time event-handling (100 milliseconds to 2 seconds - the time taken to make an action since data delivery) [3].

Two technologies are enough to use logical data marts in the Hadoop ecosystem - one for data storage, and the second - for data analytics (a tool with which we will access the data mart and “ask questions”). The technology for storing data is a file system called HDFS (Hadoop Distributed File System), the second technology used for data analytics is Apache Spark.

Both of these technologies use the principles of distribution and data parallelism. The data parallelism model can be considered to address the big data challenges, it means that large-scale data may be broken into little subsets, which further can be processed independently from each other. Each subset may have various numbers of data samples and each subset may or may not include redundant data samples. Then, a particular data analysis algorithm is performed independently in local machines (or computer nodes) and conducted over each little subset. Finally, the results of processing of each subset are combined to generate the final result through a single combination part.

The distributed method differs from the traditional one, which explicitly executes the data mining algorithm over a dataset on a single computer. As we can assume, the computation time for

the traditional single-machine approach increases greatly when the dataset size becomes very big, but the distributed approach can easily solve this large-scale dataset issue [4].

The Hadoop Distributed File System (HDFS) enables data to be stored on a large number of connected storage devices in an easily accessible format. In this way, we can say that a single logical data mart is stored as a set of split parts in a distributed file system. A "file system" is a mechanism that a computer uses to store information so that it can be found and used.

The HDFS has an architecture including master and slave. The HDFS cluster has a name-node, which is the main server (master) that controls the file system spaces and clients' access to files. In addition to the name-node, there is a variety of data-nodes (slaves) that handle the storage connected to the nodes they operate on, usually one per node in the cluster. The HDFS creates some separate logical spaces in a file system and enables user data storage in folders. A file is broken into one or more blocks, and these blocks make up a data-node package. Operations over the file system (opening, closing, and renaming files and folders) are performed by the name-node. The data-nodes are used for processing read and write operations over the files on the clients' request.

In addition to the technology for storing a logical data mart, it is equally important to have a technology which provides the functionality to use this logical data mart and perform analytics. However, there is a problem related to the distributed manner of storing files in the HDFS – one of the big data problems. It is the need to manually partition the chosen data and assign separate computer nodes (or processors) to execute the analytical task over the partitioned subsets. To deal with this problem, Spark technology has been invented. For parallel processing purposes, this can be called a new generation of computing methods.

Apache Spark is a general-purpose cluster-computing system distributed through open-source. With implicit data parallelism and fault tolerance, Spark provides an architecture for programming whole clusters originally generated at the University of California, AMPLab of Berkeley [5].

To use Spark, developers write a driver program that executes their application's high-level control flow and launches multiple parallel operations in RAM, which greatly increases the speed of computational operations. Spark has the functionality that allows us to read distributed logical data marts from HDFS by creating dataframes. A dataframe in Spark is a distributed data set grouped into designated columns. Conceptually, it is like a table in a relational database. In a usual sense, DataFrame is a copy of a logical data mart in the HDFS, which is used for performing SQL-like operations over the data and making some analytics, i.e., it is used for "asking questions" to a logical data mart. Moreover, Spark is used for processing big amounts of raw data and creating logical data marts from these data with their further storage in the Hadoop Distributed File System.

**Conclusion**

To summarize all the information presented, the demonstrated logical data mart is a convenient and optimal tool in the Big Data ecosystem that combines the functional power of business intelligence systems with the ability to process massive amounts of data. The article illustrates the main features of a logical data mart, which is used to aggregate data from various sources into a single entity. The Hadoop technology has grown in popularity because of its benefits such as rapid data processing for creating logical data marts and storing of large volumes of data, demonstrating high performance of the experiment conducted in this study.

REFERENCES

1. Golov N., Rönnbäck L. Big Data normalization for massively parallel processing databases // Computer Standards & Interfaces. - 2017. - Vol. 54. - P. 86-93.
2. Raevich, A.P., Dobronets, B.S. Developing a conceptual model of operational-analytical data marts // Modelling, optimization and information technologies. - 2019. - Vol.7 - No.4 - P. 1-13.
3. Ramesh Hari. How to build a data warehouse on Hadoop [Electronic resource] URL: http://etlcode.blogspot.com/2016/08/how-to-build-datawarehouse-on-hadoop.html (accessed: 05.05.2021)

4. Tsai, C.-F., Lin, W.-C., Ke, S.-W. Big data mining with parallel computing: A comparison of distributed and MapReduce methodologies. Journal of Systems and Software – 2016. – P. 2–3.
5. Zaharia M., Chowdhury M., Franklin M.J., Shenker S., Stoica I. Spark: Cluster Computing with Working Sets – 2010. – P. 3-5.

**Сагадиев Р.Т., Шайкемелев Г.Т.**

**Представление логической витрины данных в экосистеме Hadoop**

**Аннотация.** В данной статье раскрывается понятие логической витрины данных внутри экосистемы больших данных Hadoop, которая применяется в производстве для аналитических целей и принятия бизнес-решения. В статье представлен углубленный взгляд на современный подход обработки больших данных, а также наглядно продемонстрировано применение логических витрин данных с использованием схем и изображений. В завершении представлен вывод о выполненной работе, включающий в себя совокупность основных идей статьи.

**Ключевые слова**: технология, логическая витрина данных, хранилище данных, обработка информации, преимущества, Hadoop.

**Сагадиев Р.Т., Шайкемелев Г.Т.**

**Hadoop экожүйесінде логикалық деректер кесіндісін ұсыну**

**Аңдатпа.** Бұл мақалада Hadoop үлкен деректер экожүйесінің логикалық деректер кесіндісінің тұжырымдамасы ашылады, ол өндіріс барысында аналитикалық мақсатта және іскери шешімдер қабылдау үшін қолданылады. Мақалада үлкен деректерді өңдеудің заманауи тәсілі терең қарастырылған, сонымен қатар сызбалар мен кескіндерді пайдаланып логикалық деректер кесіндісін қолдану нақты көрсетілген. Жұмыс аяқталғаннан кейін мақаланың негізгі идеяларының жиынтығымен қоса, орындалған жұмыс туралы қорытынды жасалды.

**Түйінді сөздер**: технологиялар, логикалық деректер кесіндіс, мәлімет қоймасы, ақпаратты өңдеу, артықшылықтар, Hadoop.

**Авторлар туралы мәлімет:**

**Сагадиев Руслан Тимурович**, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының магистранті, Халықаралық ақпараттық технологиялар университеті.

**Шайкемелев Галымжан Тимурович**, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының магистранті, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Сагадиев Руслан Тимурович,** магистрант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**Шайкемелев Галымжан Тимурович,** магистрант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**About the authors:**

**Ruslan T. Sagadiyev,** master student, Department of Computer Engineering and Information Security, International Information Technology University.

**Galymzhan T. Shaikemelev,** master student, Department of Computer Engineering and Information Security, International Information Technology University.

УДК 004.056.53

**Бейсенбек Е.Б.[1], Дузбаев Н.Т.[2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан,

# СОВРЕМЕННЫЕ СПОСОБЫ ВЗЛОМА И ЗАЩИТЫ ПО

**Аннотация.** В статье представлены основные отличия лицензионного продукта от контрафактной копии, также были разобраны способы и методы обороны продукта, помимо этого представлен список инструментов, применяемых взломщиками (хакерами) для взлома программ, с краткими описаниями. Кроме того, в статье вы можете найти некоторые из основных методов взлома продукта. В процессе исследования внимание было обращено на статистику применения пиратского продукта в странах СНГ, предоставленную компанией BSA (Business Software Alliance), основанной в 1988 году, которая представляет интересы крупнейших в мире разработчиков программного обеспечения и ведет борьбу с нарушениями авторских прав на программное обеспечение, создаваемое членами этой организации. Итогом исследования стали рекомендации по использованию продуктов для защиты ПО с описанием каждого из них.

**Ключевые слова**: взлом, киберзащита, лицензия, разработка ПО, способы защиты данных, анализ рынка программного обеспечения.

## Введение

По данным BSA [1], доля применения пиратского программного обеспечения в СНГ достигает 85%. Это значит, что практически 9 из 10 копий продукта лишают производителя выгоды. Для стран Европы и Америки эта цифра намного ниже: там распространение контрафактных копий тоже лишает компании ощутимой части прибыли, но это не приводит к кризису в их бизнесе. Это происходит из-за того, что большинство западных компаний очень давно на этом рынке, отчасти поэтому они привыкли к такого рода убыткам, а пользователи в большинстве случаев предпочитают официальный продукт, предоставляемый самой компанией. В СНГ пользователи предпочитают «бесплатный» или дешевый продукт, не обращая внимания на то, что покупают: официальный продукт или пиратскую версию. В результате компании терпят колоссальные убытки, а маленькие и средние студии просто не выживают в таких условиях.

С пиратством можно бороться разными способами. Самый распространенный из них — это защита правомерных методов. Это значит, что такого рода нарушения авторских прав должны быть правильно описаны в законе, и государство должно в рамках этих законов преследовать взломщиков и привлекать их к ответственности. Но если взять в пример Казахстан, то у нашего государства есть достаточно других проблем, и до таких деталей правоохранительным органам пока нет дела.

Один из самых популярных способов борьбы с пиратством — это экономический метод. Он заключается в том, что цена продукта опускается так низко, что становится близкой к цене пиратского продукта. В основном, если пользователю предложить два продукта в виде пиратской и официальной версии примерно в одинаковом ценовом сегменте, то он предпочтет приобрести официальный продукт вместо пиратской копии. Несмотря на это, экономическая война не всем производителям продуктов подходит одинаково, так как себестоимость продукта часто в несколько раз превышает цену, за которую продаются пиратские версии, и для производителя нет смысла продавать продукт в убыток себе. В таких случаях производители в основном используют продукты других компаний, которые нацелены на защиту от любого рода взлома и нелегального копирования. Эффективная защита как правило намного уменьшает риск потери прибыли компании при атаке взломщиков.

На графике ниже (рис. 1) описаны объемы продаж хорошо защищенного и плохо защищенного продукта. Очевидно, что, если продукт не защищен должным образом, пираты с легкостью обойдут защиту и будут продавать продукт дешевле, чем производитель, и в результате будут лидировать по продажам, тем самым уменьшая прибыль производителя. В случае, если продукт защищен как положено, то пиратам будет непросто взломать его. Это даст производителю время на то, чтобы получить намеченную прибыль и продолжать дальше оставаться на рынке.

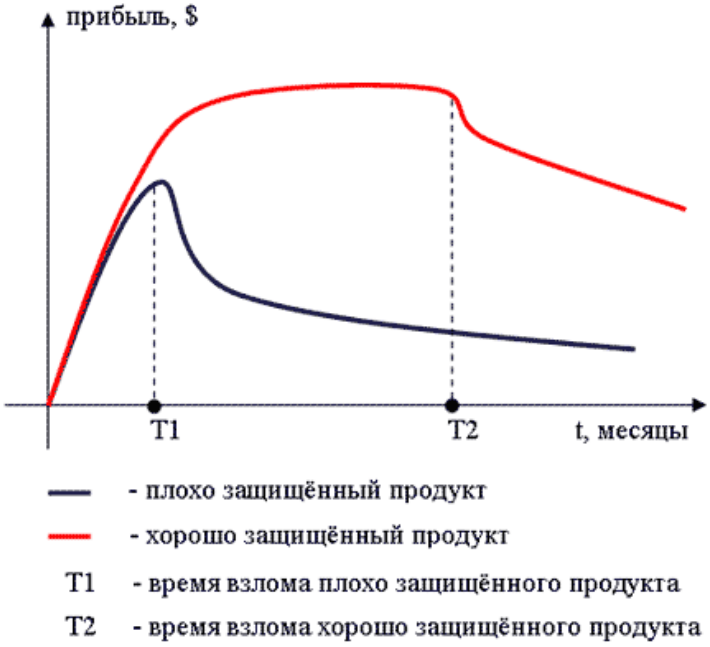



*Рисунок 1 - Динамическая зависимость прибыли от степени защищенности продукта*

Способы и методы защиты постоянно развиваются, их количество увеличивается, а срок взлома продукта в большинстве случаев занимает от недели до шести месяцев. Имелись в истории случаи, когда взламывали еще не вышедшую в продажу игру и распространяли ее до того, как начинались продажи от производителя. Ниже пойдет речь о том, как избежать этого.

Перейдем к обсуждению передовых средств обороны ПО.

Перечень основных систем защиты:

– Cd-Cops;
– LaserLock;
– StarForce;
– SafeDisk;
– SecuRom;
– TAGES;
– Denuvo Anti-Tamper [3].

Продукты «Cd-Cops» и «TAGES» готовы полноценно противостоять копированию, но абсолютно не в состоянии противостоять дизассемблерам, дамперам и отладчикам.

Самая надежная защита из списка выше — это «Denuvo Anti-Tamper». Она использует решения, делающие ее взлом практически невозможным. Принцип работы Denuvo достаточно сложен: защита циклично переносит файлы продукта в оперативную память с жесткого диска и обратно. Операция повторяется несколько раз в течение минуты. За час работы Denuvo переносит часть кода туда и обратно 150000 раз. Часть кода, которая переносится, находится в одном блоке памяти. Вследствие данных действий блок памяти подвергается опасности сгореть или выйти из строя.

Разработчики системы защиты «Denuvo Anti-Tamper» оказались не так просты, чтобы признавать, что их защита все время декодирует и зашифровывает код защищенного продукта. Они провели тесты, на которые были приглашены журналисты, чтобы их клиенты и все остальные убедились в безопасности «Denuvo Anti-Tamper» для жестких дисков типа SSD. Результаты теста показали, что их защита не наносит вреда жестким дискам. Данный тест был широко распространен в интернете.

Но пользователи продуктов с данной защитой результатов не приняли. Они самостоятельно провели независимый эксперимент, который показал, что за 40 минут использования продукта с защитой Denuvo, программа провела операций на 30 гигабайт, что очень много для накопителя за такое короткое время. Как известно, у SSD ограниченное время жизни, поэтому данные операции значительно приблизят срок его службы к концу [4].

Далее описаны параметры лицензионного продукта, предлагаемого изготовителем, в сопоставлении с контрафактной копией:

– лицензионные продукты преимущественно продаются через специальные интернет-площадки, некоторые компании продают прямо у себя на сайте;

– на упаковке поддельных дисков отсутствуют голографические наклейки;

– в случае, если на купленном диске есть метки производителя, но невысокого качества (потертости, неполная печать, другой цвет печати и.т.д.), то этот диск — нелегальная копия [2];

– стоимость лицензионной продукции на порядок выше;

– ключевое отличие пиратской копии в том, что на обратной стороне диска отсутствуют название организации-изготовителя и номера лицензии [5].

Приведем список инструментов, которые наиболее часто используются взломщиками. Их можно объединить в несколько групп: отладчики, дизассемблеры, средства слежки за действиями программы, средства пассивного анализа, прочие утилиты [6].

Способы взлома продукта: побитовое копирование, эмулирование, отладчики, дизассемблеры и дамперы.

Побитовое копирование — это копирование диска для распространения копии.

Метод эмулирования (англ. emulation) используется, когда хакер знает способ получения не копируемой метки на диске (метод получения информации о нестабильном участке диска), а метод дизассемблирования — не самый удобный способ получения данных о программе.

Следующий этап взлома — это использование таких инструментов, как отладчик, дизассемблер и дампер, которые применяются в том случае, если скопировать продукт не получилось, и о защите продукта все еще ничего неизвестно. Инструмент отладка применяется для запуска программы «пошагово» с целью найти брешь в защите, для этого используются специальные программы, называемые «отладчиком». Дизассемблирование — это перевод программы на язык программирования «Ассемблер». Дампер — более продвинутый способ анализа ПО, где транслируется содержимое оперативной памяти на момент начала исполнения программы.

**Заключение**

Лицензирование ПО и его взлом — это схватка между разработчиками программного обеспечения и хакерами-взломщиками, иными словами пиратами, которая вряд ли будет завершена в скором времени. Судя по статистике, описанной в самом начале статьи, можно точно сказать, что сторона разработчиков терпит большие убытки из-за активных нападений стороны взломщиков. Но стоит отметить, что на некоторых участках фронта разработчики пока не видели поражения. Поэтому, если ваша цель — защита продукта от действий злоумышленников, то следует обратить внимание на такие продукты, как «Denuvo Anti-Tamper» и «StarForce». К ним прилагаются собственный SDK, который упрощает процесс внедрения защиты в продукт, а также все инструкции и рекомендации в деталях.

Несмотря на все это, даже использование еще не взломанного средства для защиты продукта не значит, что не найдутся новые способы и средства взлома.

СПИСОК ЛИТЕРАТУРЫ

1. Статистика лицензионных ПО. [Электронный ресурс] URL: https://www.bsa.org/reports (дата обращения: 15.03.2021)
2. Как отличить лицензионный диск (CD, DVD) от пиратского? [Электронный ресурс] URL: http://www.genon.ru/GetAnswer.aspx?qid=5513ecfD-dc4f-4cda-b42e-168d4508e44d (дата обращения: 17.03.2021)
3. Анализ средств защиты компакт-дисков от несанкционированного копирования. [Электронный ресурс] URL: http://otherreferats.allbest.ru/programming/00238674_0.html (дата обращения: 22.03.2021)
4. Что такое Denuvo? [Электронный ресурс] URL: http://igrotop.com/posts/3682#:~:text=%D0%9F%D1%80%D0%B8%D0%BD%D1%86%D0%B8%D0%BF%20%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D1%8B%20Denuvo%20%D0%B4%D0%BE%D1%81%D1%82%D0%B0%D1%82%D0%BE%D1%87%D0%BD%D0%BE%20%D1%81%D0%BB%D0%BE%D0%B6%D0%B5%D0%BD,%D1%82%D1%83%D0%B4%D0%B0%2D%D1%81%D1%8E%D0%B4%D0%B0%20150%20%D1%82%D1%8B%D1%81%D1%8F%D1%87%20%D1%80%D0%B0%D0%B7 (дата обращения: 23.03.2021)
5. Совет, как отличить лицензионный диск от пиратского. [Электронный ресурс] URL: http://sowetu.ru/read/4090.html (дата обращения: 26.03.2021)
6. Защита программ от взлома. [Электронный ресурс] URL: http://z-oleg.com/secur/articles/progprotect.php (дата обращения: 26.03.2021)

REFERENCES

1. Statistics of licensed software [Electronic resource] URL: https://www.bsa.org/reports (date of the application: 15.03.2021)
2. How to distinguish a licensed disc (CD, DVD) from a pirated one [Electronic resource] URL: http://www.genon.ru/GetAnswer.aspx?qid=5513ecfD-dc4f-4cda-b42e-168d4508e44d (date of the application: 17.03.2021)
3. Analysis of means of protecting CDs from unauthorized copying [Electronic resource] URL: http://otherreferats.allbest.ru/programming/00238674_0.html (date of the application: 22.03.2021)
4. What is Denuvo [Electronic resource] URL: http://igrotop.com/posts/3682#:~:text=%D0%9F%D1%80%D0%B8%D0%BD%D1%86%D0%B8%D0%BF%20%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D1%8B%20Denuvo%20%D0%B4%D0%BE%D1%81%D1%82%D0%B0%D1%82%D0%BE%D1%87%D0%BD%D0%BE%20%D1%81%D0%BB%D0%BE%D0%B6%D0%B5%D0%BD,%D1%82%D1%83%D0%B4%D0%B0%2D%D1%81%D1%8E%D0%B4%D0%B0%20150%20%D1%82%D1%8B%D1%81%D1%8F%D1%87%20%D1%80%D0%B0%D0%B7 (date of the application: 23.03.2021)
5. Advice on how to distinguish a licensed disc from a pirated one [Electronic resource] URL: http://sowetu.ru/read/4090.html (date of the application: 26.03.2021)
6. Protection of programs from hacking [Electronic resource] URL: http://z-oleg.com/secur/articles/progprotect.php (date of the application: 26.03.2021)

**Бейсенбек Е.Б., Дузбаев Н.Т.**

**Бағдарламалық жасақтаманы бұзудың және қорғаудың заманауи әдістері**

**Аңдатпа**. Мақалада лицензияланған өнім мен контрафактілік көшірме арасындағы негізгі айырмашылықтар келтірілген, сонымен қатар өнімді қорғау әдістері мен тәсілдері талданған, бұған қоса, крекерлер (хакерлер) бағдарламаларды бұзу үшін пайдаланатын қысқаша сипаттамалары бар құралдар тізімі берілген. Бұған қоса, мақалада өнімді бұзудың кейбір негізгі әдістерін таба аласыз. Зерттеу барысында әлемдегі ең ірі бағдарламалық жасақтама жасаушылардың мүдделерін білдіретін және авторлық құқықпен күресетін 1988 жылы құрылған BSA (Business Software Alliance) ұсынған ТМД елдері арасында қарақшылық өнімді пайдалану статистикасы ескерілді, осы ұйым мүшелері жасаған бағдарламалық жасақтаманың

бұзылуы. Зерттеу нәтижесінде бағдарламалық жасақтаманы қорғау құралдарын пайдалану және оларды танып, білу туралы кең ұсыныстар берілді.

**Түйінді сөздер**: хакерлік, кибер қорғаныс, лицензия, бағдарламалық жасақтама жасау, деректерді қорғау әдістері, бағдарламалық жасақтама нарығын талдау.

**Beisenbek Y.B., Duzbaev N.T.**

**Modern methods of hacking and protection software**

**Abstract.** The article delineates the main differences between a licensed product and a counterfeit copy, also analyzes the ways and methods of protecting the product. Additionally, it presents a list of tools with brief descriptions used by crackers (hackers) to crack programs and some of the basic methods of hacking a product. The research hinges on the statistics on the use of pirated products among the CIS countries provided by BSA (Business Software Alliance), set up in 1988, which represents the interests of the world's largest software developers and fights against copyright infringement on the software created by its members. As a result, the study presents some comments on the nature of the analyzed software protection products and suggests recommendations on their use.

**Keywords:** hacking, cyber protection, license, software development, data protection methods, software market analysis.

**Авторлар туралы мәлімет:**

**Бейсенбек Ерболат Бақытжанұлы**, Халықаралық ақпараттық технологиялар университетінің магистрі.

**Дузбаев Нуржан Токкужаевич**, PhD, доцент, Халықаралық ақпараттық технологиялар университеті, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының проректоры.

**Сведения об авторах:**

**Бейсенбек Ерболат Бакытжанулы,** магистрант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**Дузбаев Нуржан Токкужаевич,** PhD, ассоциированный профессор, проректор по цифровизации и инновациям, Международный университет информационных технологий.

**About the authors:**

**Beisenbek Y. Bakhyzhanuly**, master student, Department of Computer Engineering and Information Security, International Information Technology University.

**Duzbaev N. Tokkuzhaevich** PhD, Vice-Rector for Digitalization and Innovation, Associate Professor, Department of Computer Engineering and Information Security, International Information Technology University.

УДК 004.42, 519.85

**Найзабаева Л.К [1], Алашыбаев Б.А [2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан

# РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА ДЛЯ ОНЛАЙН-МАГАЗИНОВ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

**Аннотация**. Среди множества последних трендов интернет-маркетинга можно отметить рекомендательные системы. Рекомендательные системы — особые приложения, направленные на прогнозирование интересов и потребностей вероятных покупателей интернет-магазинов, являющиеся комфортным инструментом выбора при приобретении продуктов и предложений в онлайн-магазинах. Принципально важными факторами, влияющими на развитие рекомендательных сервисов, являются польза и удобство одновременно и для потребителя, и для интернет-магазина. Пользователь, прежде всего, получает удобство интуитивного выбора. Для магазина открываются такие возможности, как увеличение среднего чека и выручки компании, альтернативная навигация во всем множестве товаров и получение информации о клиентах. Современные рекомендательные сервисы повышают наполненность онлайн-корзин на 12–60%, что обычно зависит от профильной направленности продукции.

**Ключевые слова:** коллаборативная фильтрация, гибридные системы, взвешенный наклон один (Weighted Slope One), модель Байеса, кластерная модель.

## Введение

Мы живем в мире информационного обилия. Выбор — это то, с чем встречается любой человек каждый день. При наличии всевозможных вероятностей в обязательном порядке появляется проблема выбора одной из альтернатив: какой продукт приобрести, какой кинофильм посмотреть, какое музыкальное произведение послушать, какую заметку прочитать и т. д. Рекомендательные системы — это комплект программных средств и способов, призванных посодействовать определенному пользователю в выборе между разных объектов [1].

Область использования рекомендательных систем многообразна. Наибольшее значение они имеют для интернет-магазинов. Персонализация онлайн-маркетинга считается отличительным трендом последних десятков лет [2]. По оценкам McKinsey, 35% дохода у Amazon и 75% у Netflix приходится как раз на рекомендуемые продукты, и данный процент, скорее всего, будет расти. Активно применяются рекомендательные системы и на веб-сайтах, производящих или публикующих контент [3].

Большинство клиентов интернет-магазинов, скорее всего, в будущем столкнется с использованием подобной программы, которая может дать нам личный совет и помочь в принятии решений [4]. Поиски новых методов создания подобной программы являются ключевой задачей той области науки, которая занимается разработкой рекомендательных систем. А задача самой системы — гарантировать сообществу пользователей предоставление доступной информации и качественных советов [5].

В связи с повсеместным внедрением онлайна и наличием большого объема статистических данных о предпочтениях пользователей научные работники всего мира стали активно заниматься разработкой этих алгоритмов в последние двадцать-тридцать лет [6]. А вследствие широкой применимости эти программы оказались достаточно эффективными.

В настоящей статье говорится о ведущих рекомендательных методах, а также приводятся примеры самых свежих разработок в данной области. Цель нашей работы — сделать обзор ведущих рекомендательных алгоритмов, рассказать о принципах способа

коллаборативной фильтрации, о системах фильтрации на базе содержания и на базе информации о пользователях, познакомить читателя с достоинствами и недостатками каждого из способов. Наиболее важным при знакомстве с данными разработками является понимание того, как их чаще всего используют, ведь все новейшие разработки в данной области направлены на улучшение именно этого параметра [7].

Для создания рекомендательных систем существует несколько методов, которые в основном базируются на коллаборативной фильтрации:

- Фильтрация, основанная на схожести пользователей;
- Фильтрация, основанная на схожести объектов (предметов);
- Фильтрация, основанная на модели;
- Модель Байеса;
- Регрессионная модель;
- Кластерная модель;
- Также достаточно распространены методы, основанные на факторизации;
- Неотрицательная матричная факторизация (NMF);
- Сингулярное разложение (SVD).

**Коллаборативная фильтрация**

Коллаборативная фильтрация — это способ, позволяющий предугадывать неизвестные предпочтения пользователя на базе популярных оценок и/или поведения иных пользователей (Segaran 2007). Данный способ базируется на представлении о том, что пользователи, идентично оценившие одни предметы системы, имеют предрасположенность идентично оценивать и иные предметы системы. Еще одно допущение состоит в том, что пользователи дают схожие оценки предмету системы по парциальной шкале, к примеру, оценивая кинофильм от одной до десяти звезд на веб-сайте imdb.com (рис. 1). Данное допущение является важным, однако не во всех системах есть возможность собирать оценки пользователей. В этих случаях прибегают к скрытому сбору информации и оценке поведения, к примеру, записывая просмотренные ролики на YouTube и рекомендуя связанные с ними материалы [8].


*Рисунок 1 - Пример оценки фильма на сайте imdb.com*

Коллаборативная фильтрация обычно делится на два подхода. Первым, и самым распространенным, считается подход, базирующийся на близости (сходстве) пользователей. Его сущность заключается в анализе прошлых оценок или поведения пользователя, исследование других пользователей, имеющих схожую “историю” и составление прогноза для неизвестных оценок [10].

В обычной ситуации образуется матрица «пользователи-предметы», смысл которой заключается в оценках определенным пользователем определенного предмета. Те ячейки, в которых нет значений, считаются неизвестными, то есть заявленному предмету пользователь не выставил оценку, значит, вероятнее всего, не воспользовался этим предметом (табл. 1).

Таблица – 1 Пример матрицы «пользователи-предметы»

| | “Даллаский клуб покупателей” | “Человек-паук” | “Волк с Уолл-стрит” | “12 лет рабства” | “Free-to-play” |
|---|---|---|---|---|---|
| Пользователь_1n | 5 | 3 | 5 | - | - |
| Пользователь_2n | 4 | - | 4 | - | 5 |
| Пользователь_3n | - | 1 | 5 | 2 | - |

Для определения близости пользователей используется некоторое количество различных алгоритмов, таких как:

- Манхэттенское расстояние;
- Евклидово расстояние;
- Коэффициент корреляции Пирсона.

Манхэттенское расстояние, или расстояние городских кварталов считается одним из базисных способов вычисления дистанции между 2-мя точками (1):

$$d(user1, user2) = \sum_{k=1}^{n} |user1_k - user2_k| \quad (1)$$

*где*

*user1, user2 – пользователи и их оценки;*

*n – количество предметов в матрице.*

Данный подход является недостаточно точным при небольшом заполнении матрицы, но его преимущество в простоте и быстрой скорости выполнения.

Евклидово расстояние содержит квадратные корни (теорема Пифагора) и рассчитывается с помощью соответствующей формулы (2):

$$d(user1, user2) = \sqrt{\sum_{k=1}^{n} (|user1_k - user2_k|)^2} \quad (2)$$

*где*

*user1, user2 – пользователи и их оценки;*

*n – количество предметов в матрице.*

Как и расстояние городских кварталов, этот способ содержит трудности при незаполненной матрице, но несложен в разработке и дешев в выполнении [10].

Более точный метод определения близости воплощен в коэффициенте корреляции Пирсона (3):

$$corr(user1, user2) = \frac{\sum_{i=0}^{n} (user1_i - \overline{user1})(user2_i - \overline{user2})}{\sqrt{\sum_{i=0}^{n} (user1_i - \overline{user1})^2} \sqrt{\sum_{i=0}^{n} (user2_i - \overline{user2})^2}} \quad (3)$$

*где*
*user1, user2 – пользователи и их оценки;*
*n – количество предметов в матрице.*
$\overline{user1}$ и $\overline{user2}$ – средняя оценка пользователей

Значение *corr(user1, user2)* может быть от -1 до 1, где -1 соответствует абсолютному несовпадению пользователей, а 1 — абсолютному совпадению.

Методы, базирующиеся на сходстве пользователей, могут быть полезны — они интуитивно понятны и просты в применении [10].

**Взвешенный наклон один (Weighted Slope One)**

Одним из самых действенных алгоритмов в предметно-ориентированной коллаборативной фильтрации считается метод взвешенный наклон один (Weighted Slope One). Его сущность заключается в поиске различий оценок между парами элементов и применении данных различий для вычисления исключений [16]. Вычисление различий между предметами производится с помощью соответствующей формулы (4):

$$dev_{i,j} = \sum_{user \in S_{i,j}(X)} \frac{user_i - user_j}{both(S_{i,j}(X))} \quad (4)$$

*где*
$both(S_{i,j}(X))$ *– число пользователей, оценивших и i-й, и j-й элемент;*
*user – оценки пользователя;*

Прогнозирование оценки предмета рассчитывается следующим образом (5):

$$P(user_j) = \frac{\sum_{i \in S(u)-\{j\}} (dev_{j,i} + u_i) card(S_{j,i}(X))}{\sum_{i \in S(u)-\{j\}} both(S_{j,i}(X))} \quad (5)$$

*где*
$both(S_{i,j}(X))$ *– число пользователей, оценивших и i-й, и j-й элемент;*
*user – оценки пользователя;*

Взвешенный наклон один (Weighted Slope One) считается замечательным способом разработки рекомендательных систем. Имея невысокие требования к памяти и большую скорость работы, он демонстрирует высокую эффективность при наличии большого количества пользователей. Однако, сложной задачей для этого метода является проблема холодного старта, которая, в общем, относится ко всем предметно-ориентированным методам коллаборативной фильтрации [11]. Род алгоритмов Scope One применяется в некоторых популярных сервисах, таких как hit flip, вебсайт рекомендаций DVD и Value Investing News, новостной портал фондовых бирж.

Второй большой группой методов коллаборативной фильтрации является фильтрация, базирующаяся на модели. Рассмотрим некоторые из них [11].

**Модель Байеса**

Одним из самых популярных классификаторов является наивный байесовский классификатор. С его помощью создают рекомендации каких-то объектов. В его основе находится вероятностная модель теоремы Байеса [11]. Для работы этого алгоритма нужно сделать модель Байеса для каждого пользователя, который изучал какие-либо объекты, на основе содержания данных объектов (для кинокартин это могут быть артисты или жанры, для новостей — главные тексты и рубрики). Для нахождения более вероятной категории нужно определить относительные вероятности приспособления какого-нибудь предмета к любой категории и выбрать категорию, имеющую самую большую вероятность (6):

$$cat = \arg max P(c) \prod P(o_i|c) \quad (6)$$

**Кластерная модель**

Одним из самых популярных алгоритмов в кластерном анализе считается способ k-средних. Он используется при делении объектов или пользователей на группы — кластеры, формирующиеся по некоторым общим признакам, количество которых задается заранее. Сущность метода — в случайном выборе k-центров кластера и сокращении суммарного квадратичного отдаления пользователей или объектов от центра кластера [11]. Это рассчитывается с помощью соответствующей формулы (7):

$$d = \sum_i^k \sum_{x_j \in K} (x_j - u_i)^2 \quad (7)$$

*где*

*k – количество векторов;*

*u – центр масс векторов из множества кластеров K.*

**Факторизация матриц**

Способы коллаборативной фильтрации довольно наглядны и просты, но с помощью факторизации матриц иногда можно получить гораздо больше результатов, например, данный способ позволяет обнаружить некоторые скрытые моменты, объединяющие объекты и пользователей. Но нужно помнить о том, что это математический способ, и для получения адекватных результатов потребуется настройка данного метода [11].

Целью неотрицательной факторизации матриц считается разложение матрицы на произведение 2-ух других матриц. В случае рекомендательной системы начальная матрица будет считаться матрицей «пользователи-объекты», а значения в ячейках — оценками данных пользователей всевозможных объектов. Если некоторых оценок нет, то с помощью факторизации вполне возможно получение данных недостающих оценок.

Математически это рассчитывается при помощи получения квадратичной погрешности, вычисления градиента от нее и получения значения матриц P и Q: (8)(9)

$$e_{i,j}^2 = (r_{i,j} - \sum_{k=1}^K p_{i,j} q_{i,j})^2 \quad (8)$$

*где*

*r – реальное значение исходной матрицы,*

*p и q – значения предполагаемых матриц P и Q*

$$p_{i,k}^{new} = p_{i,k} + 2\alpha e_{i,j} q_{k,j}$$
$$q_{k,j}^{new} = q_{k,j} + 2\alpha e_{i,j} p_{i,k} \quad (9)$$

*где*

$p_{i,k}^{new}$ *и* $q_{k,j}^{new}$ *– новые значения в матрицах P и Q*

$2\alpha e$ *– градиент из квадратичной ошибки*

Этот метод требует большого числа итераций (>5000) для получения верных итогов.

**Проблемы существующих подходов**

• *Недостаток информации о пользователях*

Основная масса пользователей различных сервисов не любит высказывать свое мнение о каких-либо предметах и выставлять им оценки, из-за чего точность прогноза утрачивается.

• *Новые пользователи*

Данная проблема также является серьезной — новый пользователь не предоставляет достаточно информации для создания качественных рекомендаций для него, поэтому рекомендации формируются или из известного в целом контента, или не формируются вообще.

• *Поддельные оценки*

На многих популярных порталах есть определенные пользователи, которые осознанно завышают или занижают оценки всем объектам системы, продвигая те или иные продукты, что мешает работать рекомендационным алгоритмам.

• *Масштабируемость*

При увеличении количества пользователей и объектов системы возрастает время выполнения алгоритмов для каждого пользователя. Отчасти это решается улучшением алгоритмов (например, для метода weighted slope one нужно только наблюдать за отклонениями между парами элементов и общим количеством элементов).

В итоге можно сказать, что все существующие методы имеют как плюсы, так и минусы. Одни методы выигрывают по временным затратам, другие — по требованиям к памяти и точности, но для данной работы было принято решение выбрать способы, базирующиеся на статистических вычислениях. Создание классификатора не всегда приводит к нужным результатам. Например, важные тексты не всегда попадают в анонсы, а ведущие новостей не дают теги к объектам новости. Оценивание каждого новостного контента, прочтенного пользователем, не всегда дает верное заключение о его предпочтениях.

Поэтому для создания рекомендательной системы было принято решение выбрать метод поиска схожих людей через взвешенный наклон один (Weighted Slope One).

**Постановка задачи**

Сформулируем задачу. Допустим, у нас будет большое количество пользователей U = $\{u_1, u_2, \dots, u_n\}$, большое количество объектов P = $\{p_1, p_2, \dots, p_m\}$ и матрица рейтингов R = $(r_{i,j})$ размера n × m, где i ∈ 1. . . n, j ∈ 1. . . m. Вероятными значениями рейтингов могут быть числа от 1 (не нравится) до 5 (нравится). Если пользователь i не оценил элемент j, то на месте $r_{i,j}$ значение будет пустующим. Описанную матрицу рейтингов можно представить в виде таблицы (табл. 2). Обозначим через $r'_{i,j}$ наш прогноз относительно того, какую оценку пользователь i поставит продукту j. Наша задача предсказать, какие оценки $r_{i,j}$ должны стоять на месте пропусков в матрице, то есть рассчитать $r'_{i,j}$.

Таблица – 2 Пример матрицы «пользователи-продукты»

| Продукты | Продукт 1 | Продукт 2 | Продукт 3 | Продукт 4 | Продукт 5 | Продукт 6 |
|---|---|---|---|---|---|---|
| Пользователь 1 | 2 | 4 | 5 | 3 | 4 | ? |
| Пользователь 2 | 1 | ? | 1 | 4 | 5 | ? |
| Пользователь 3 | 3 | 4 | ? | 1 | 4 | ? |
| Пользователь 4 | ? | ? | 4 | 2 | ? | 2 |
| Пользователь 5 | ? | 4 | 5 | 3 | ? | ? |

Затем для каждого пользователя u на основе спрогнозированных оценок $r'_{i,j}$, нам нужно сформировать список из N продуктов, которые соответствуют предпочтениям пользователя и которые еще им не оценены. Список этих N продуктов обозначим через N-мерный вектор ( $p_{i_1}, p_{i_2}, \dots, p_{i_n}$). Таким образом, математическая задача звучит так:

**Дано:**

U = $\{u_1, u_2, \dots, u_n\}$ — множество пользователей,

P = $\{p_1, p_2, \dots, p_m\}$ — множество продуктов,

R = ($r_{i,j}$) - матрица рейтингов размера n × m, где на месте $r_{i,j}$ будет стоять число, если пользователь $u_i$ оценил продукт $p_j$ и пусто в ином случае. N — требуемое число рекомендаций, которые нужно получить от системы.

Требуется найти: Для данного пользователя $u_i$ найти N-мерный вектор ( $p_{i_1}, p_{i_2}, \dots, p_{i_n}$ ), где продукты $p_{i_k}$, k ∈ N еще не оценены этим пользователем, то есть в матрице рейтингов R = ($r_{i,j}$ ) пусто на месте $r'_{i,i_k}$, а также проследить, чтобы эти продукты наиболее точно удовлетворяли предпочтения пользователя, то есть прогнозные рейтинги $r'_{i,i_k}$ были наибольшими.

**Решение:**

Взвешенный наклон один (Weighted Slope One)

Вычисление различий между продуктами производится с помощью соответствующей формулы (10):

$$dev_{i_k,j} = \frac{1}{|U_{i_k,j}|} \sum_{u \in \{U_{i_k,j}\}} (r_{u,i_k} - r_{u,j}) \quad (10)$$

*где*

$|U_{i_k,j}|$ — *число пользователей, оценивших и* $i_k$*-й, и j-й элемент;*

$\{U_{i_k,j}\}$ — *набор пользователей, оценивших и* $i_k$*-й, и j-й элемент;*

*u — оценки пользователя.*

Прогноз оценки предмета рассчитывается следующим образом (11):

$$r'_{i,i_k} = \frac{1}{|S_u - \{i_k\}|} \sum_{j \in \{S_u - \{i_k\}\}} (dev_{i_k,j} + r_{i,j}) \quad (11)$$

*где*

$\{S_u - \{i_k\}\}$ — *все элементы, оцененные пользователем u, кроме элемента* $i_k$

$|S_u - \{i_k\}|$ — *представляет количество элементов, оцененных пользователем u, кроме элемента* $i_k$

$r_{i,j}$ — *рейтинг пользователя u для элемента j.*

Итак, спрогнозируем неоцененные товары пользователем $u_i$.

После создания прогноза необходимо составить список Top N (N-мерный вектор) для рекомендаций.

Итого, есть все прогнозы для всех продуктов, которые пользователь $u_i$ еще не оценил.

Теперь составим список Top N (N-мерный вектор):

1. Пусть рейтинги будут от 1 до 5.
2. Если прогноз больше 5, мы посчитаем, что это 5.
3. Нужен лимит рекомендаций, поставим 3.5.

После этого добавим в список Top N (N-мерный вектор) все продукты с прогнозом от 3,5 до 5 и будем рекомендовать этот список с продуктами для пользователя $u_i$.

**Заключение**

В настоящей работе рассмотрены главные разновидности рекомендательных систем и основы их создания. Приведен детальный перечень алгоритмов коллаборативной фильтрации, показаны методы оценки свойств схожих объектов.

В качестве основного способа решения задачи выбран способ, основанный на коллаборативной фильтрации и реализованы методы, базирующиеся как на близости пользователей и объектов, так и на факторизации исходных данных. Наилучший результат продемонстрировал метод взвешенный наклон один (Weighted Slope One).

Работа может быть продолжена путем совершенствования базисных алгоритмов, проведения экспериментов с построением других гибридных моделей, применения

добавочных метаданных для создания систем, основанных на информации о пользователе, и систем на базе контента.

СПИСОК ЛИТЕРАТУРЫ


1. М.А.Терехин, Разработка системы рекомендаций. [Электронный ресурс] URL: https://elib.bsu.by/bitstream/123456789/232575/1/118–121.pdf (accessed: 2019)
2. Mackenzie I., Meyer C., Noble S. How retailers can keep up with consumers. [Электронный ресурс] URL: https://www.mckinsey.com/industries/retail/our-insights/howretailers-can-keep-up-with-consumers (дата обращения: 26.03.2021)
3. Daniel Lemire, Anna Maclachlan. «Slope One Predictors for Online Rating-Based Collaborative Filtering. » In SIAM Data Mining (SDM'15), 2019. [Электронный ресурс] URL: https://www.researchgate.net/publication/326079280_A_trust-based_collaborative_filtering_algorithm_for_E-commerce_recommendation_system (дата обращения: 26.03.2021)
4. Lefats'e Manamolela, Tranos Zuva, Martin Appiah. Collaborative Filtering Recommendation Systems Algorithms, Strengths and Open Issues, 2020. [Электронный ресурс] URL: https://link.springer.com/chapter/10.1007%2F978-3-030-63319-6_14 (дата обращения: 26.03.2021)
5. Pawel Matuszek, João Vinagre, Myra Spiliopoulos «Forgetting methods for incremental matrix factorization in recommender systems.» Conference: ACM SAC 2015, [Электронный ресурс] URL: https://www.researchgate.net/publication/280246432_Forgetting_methods_for_incremental_matrix_factorization_in_recommender_systems (дата обращения: 26.03.2021)
6. Sanderson, Dan. Programming Google App Engine. O'Reilly Media, 2019. [Электронный ресурс] URL: https://pdfweek.com/downloads/programming%20google%20app%20engine%20programming%20gamenetore%20pdf (дата обращения: 26.03.2021)
7. Segaran, Toby. Programming Collective Intelligence. O'Reilly Media, 2017. O'Reilly Media, Inc, – T. 13, No. 1. – P.117–134.
8. Hill W., Stead L., Rosenstein M., Furnas G. Recommending and Evaluating Choices in a Virtual Community of Use // Proceeding Conference Human Factors in Computing Systems, 2015. P. 194–201.
9. Resnick P., Iakovou N., Sushak M., Bergstrom P., Riedl J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews // Proceeding 1994 Computer Supported Cooperative Work Conference, 2014. P. 175–186.


**Найзабаева Л., Алашыбаев Б.А.**

**Машиналық оқытуды қолдану арқылы интернет-дүкендерге арналған ұсыныс жүйесі**

**Аңдатпа.** Интернет-маркетингтің соңғы тенденцияларының арасында ұсыныс жүйелерін бөлуге болады. Ұсынушы жүйелер – бұл интернет-дүкендердің тауарлары мен қызметтерін сатып алу кезінде таңдаудың ыңғайлы құралы болып табылатын интернет-дүкендердің әлеуетті клиенттерінің қызығушылықтары мен қажеттіліктерін болжауға бағытталған арнайы қосымшалар. Ұсыныс қызметтері пайдаланушы үшін де, интернет-дүкен үшін де пайдалы және ыңғайлы болуы өте маңызды. Ең алдымен қолданушының ыңғайлы және интуитивті таңдауы бар. Сонымен қатар дүкенге бару кезінде орташа чек пен кірісті ұлғайту, тауарлардың алуан түріндегі альтернативті навигация және тұтынушылар туралы дереккөзі ақпарат мүмкіндіктерін ашады. Бұл, әдетте өнімнің профиліне байланысты, бүгінгі таңда заманауи ұсынымдық қызметтер онлайн-дүкен арбаларының құрамын 12-60%-ға арттырады.

**Түйінді сөздер:** бірлескен сүзу, гибридті жүйелер, өлшенген көлбеу, Байес моделі, кластерлік модель.

**Naizabayeva L., Alashybayev B.A.**

**A recommendation system for online stores using Machine Learning**

**Abstract.** Recommender systems can be singled out among the latest trends in Internet marketing. Recommender systems are special applications focused on predicting the interests and needs of potential customers of online stores, which are convenient tools for choosing when buying goods and services in online stores. It is fundamentally important that recommendation services are useful and convenient for both the user and the online store. The user, first of all, has the convenience and intuitiveness of the choice. At the same time, the store opens up such opportunities as increasing the average check and revenue per visit, alternative navigation in the entire variety of products and a source of customer information. Today, modern recommendation services increase the content of online shopping carts by 12-60%, which usually depends on the profile of the product.

**Keywords:** collaborative filtration, hybrid systems, Weighted Slope One, Bayesian model, Cluster model.

**Авторлар туралы мәлімет:**

**Найзабаева Лязат**, т.ғ.д., Халықаралық ақпараттық технологиялар университеті, «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры.

**Алашыбаев Бекарыс Алашыбаевич**, «Ақпараттық жүйелер» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Найзабаева Лязат**, доктор технических наук, ассоциированный профессор кафедры «Информационные системы», Международный университет информационных технологий.

**Алашыбаев Бекарыс Алашыбаевич**, магистрант кафедры «Информационные системы», Международный университет информационных технологий.

**About the authors:**

**Lyazat Nayzabayeva,** Dr. Sc. (Technology), Associate Professor, Department of Information Systems, International Information Technology University.

**Bekarys A. Alashybayev,** master student, Department of Information Systems, International Information Technology University.

УДК 004.45, 004.41

**Meirambaiuly N.[1], Duzbaev N.T.[2]**

[1,2] International Information Technology University, Almaty, Kazakhstan

# MONITORING OF STATIONARY SOURCES OF POLLUTANT EMISSIONS IN ALMATY

**Abstract.** The article introduces a system concept for analyzing the Almaty's air basin. Air pollution monitoring plays a key role in environmental issues. However, in order to develop an environmental information system, it is necessary to conduct a careful analysis of natural events in order to determine their direct impact on the state of the environment, in this case the atmosphere. To assess the degree of pollution, emissions must be measured at specific intervals in specific locations across the settlement, depending on the study approach used. To numerically implement the data obtained, a model for determining the amount of pollutant concentration should be used, which will lead to a numerical expression representing the atmospheric conditions in a certain industrial region. The purpose of the project is to programmatically implement the above operations, namely to develop software based on a specific appropriate mathematical model that allows for the calculation of pollutant concentrations and, on that basis, the complex index of atmospheric pollution. To analyze and make specific decisions related to reduction of pollution levels, it is necessary to track the dynamics of changes in the atmospheric pollution index over long periods of time, which necessitates the creation of a database to record measurements as well as the ability to view them visually at any time, for example, on a map.

**Keywords:** monitoring, contaminated substances, atmosphere, emission, air area, industry, gray dioxide, nitrogen oxide, ecology.

## Introduction

The indicators of particular pollution and discharges of pollutants per unit of gross domestic product, which reflect the environmental efficiency of the national economy, have significantly deteriorated in the Republic of Kazakhstan over the last decade. As a result, the industrial sector of Almaty in 2010-2020 has undergone significant changes. Anti-dumping and import substitution policies were implemented, as well as measures to reorganize inefficient and idle production facilities, consequently providing support for small enterprises. The number of industrial enterprises increased 3.5 times during this period, including small ones - 4.6 times. The specificity of the environmental situation in Kazakhstan is that there is a high proportion of physically worn out and obsolete production facilities and technologies; only according to official data, the deterioration of production capacities of most industrial enterprises exceeds 70%. There is a portion of the Almaty region's industrial companies in Almaty, some of which have been redesigned for warehousing and passive manufacturing. There are factories for metalworking and mechanical engineering, fruit canning, meat canning, dairy, textile, fur, and other factories in the city. In addition, there are sewing, shoe, knitwear factories, and a cotton-spinning mill, household chemicals, building materials, industrial and household boiler facilities. The priority sectors are the construction industry, the electric power industry, the processing industry, and mechanical engineering. The main sources of pollution are CHPP-1 and CHPP-2 near Almaty's northwestern border, an asphalt and bitumen plant, AZTM, AHBK, a fruit canning plant in the city's central and western districts near Ryskulov Street and west of Seifullin. In addition, there is such a "pollutant" as a machine-tool company and an elevator located in the city's northern section. Almaty is home to more than 350 businesses, all of which have an impact on the city's air area. Furthermore, a huge number of diverse small enterprises have recently developed not only in the industrial zone but also across the

city, with a total impact of emissions comparable to that of industrial giants, thereby determining the current level of air pollution.

**Requirements for developing the system. Input and output data.**

Environmental requirements: The Atmosphere Monitoring system must run on personal PCs that are running Windows 7, Windows 8, or Windows 10. A database management system (DBMS) must be deployed on a server that is accessible to all users' PCs via common network protocols.

**Requirements for software development and life cycle:**

-conducting the trial operation of the system;
- elimination of possible problems in the system;
- further modification and improvement of the software product.

**System requirements:**

-getting weather data by location;
-storing the input data;
-viewing the stored data for a certain period of time;
- inputting the incoming data;
-calculation of the pollutants' concentration;
-calculation of the index of atmospheric pollution for one pollutant;
-calculation of a complex index of atmospheric pollution for five main pollutants;
- displaying a graph for calculating the air pollution index;
-printing a graph for calculating the air pollution index.
-visualizing it all on the map.

**The input data for this program will be the following:**

- from the very beginning, it is necessary to determine the number of stationary sources of pollution of the investigated industrial facility, since, depending on the area of the settlement, the number of sources of pollution may vary;

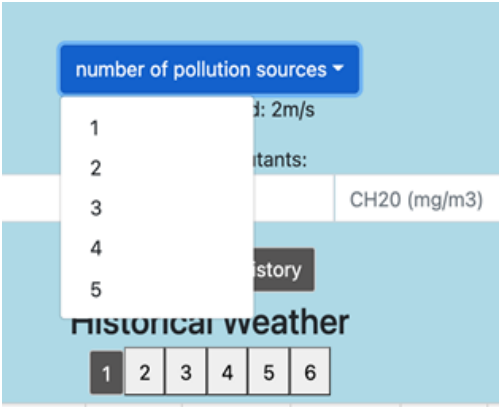



*Figure 1 - Input of the number of pollution sources*

-the initial wind speed is obtained from OpenApi services;

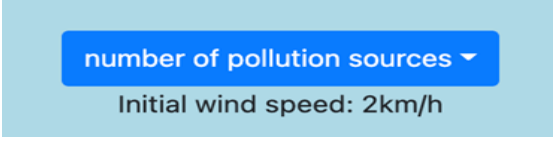


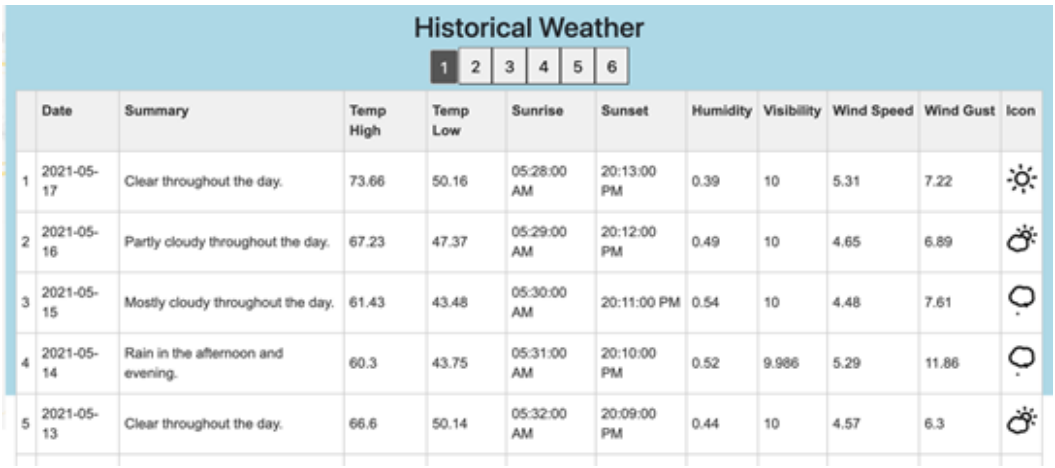

*Figure 2 - Initial wind speed*

-data on measurements of the main pollutants, which are obtained by measuring special devices (gas analyzers, aspirators);


*Figure 3-Input of pollutant measurements*

- The date and history of the weather are obtained using OpenApi services;

-




| | Date | Summary | Temp High | Temp Low | Sunrise | Sunset | Humidity | Visibility | Wind Speed | Wind Gust | Icon |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2021-05-17 | Clear throughout the day. | 73.66 | 50.16 | 05:28:00 AM | 20:13:00 PM | 0.39 | 10 | 5.31 | 7.22 | |
| 2 | 2021-05-16 | Partly cloudy throughout the day. | 67.23 | 47.37 | 05:29:00 AM | 20:12:00 PM | 0.49 | 10 | 4.65 | 6.89 | |
| 3 | 2021-05-15 | Mostly cloudy throughout the day. | 61.43 | 43.48 | 05:30:00 AM | 20:11:00 PM | 0.54 | 10 | 4.48 | 7.61 | |
| 4 | 2021-05-14 | Rain in the afternoon and evening. | 60.3 | 43.75 | 05:31:00 AM | 20:10:00 PM | 0.52 | 9.986 | 5.29 | 11.86 | |
| 5 | 2021-05-13 | Clear throughout the day. | 66.6 | 50.14 | 05:32:00 AM | 20:09:00 PM | 0.44 | 10 | 4.57 | 6.3 | |

*Figure 4-History of weather by dates*

- coordinates of the source of pollution:

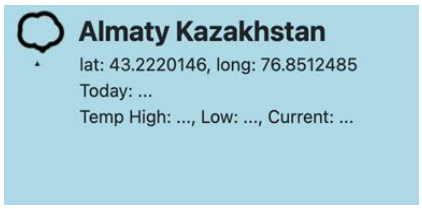



*Figure 5 - Obtaining the coordinates of the pollution source*

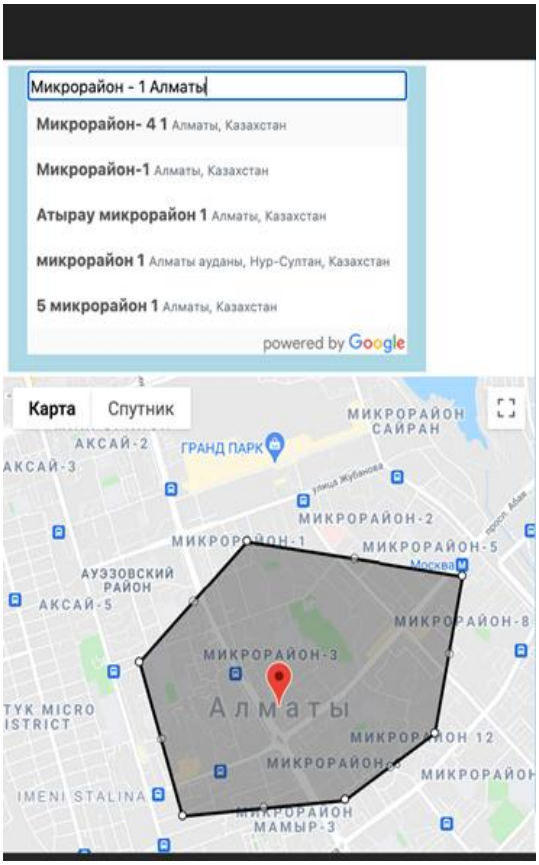



*Figure 6 - Finding a location*

The output information includes calculated values for atmospheric pollution indices for a specific substance, as well as a complex atmospheric pollution index for five major air pollutants, which can be used to draw conclusions about the state of the surface layers of a specific object under study and, as a result, take appropriate measures to reduce the pollution levels.

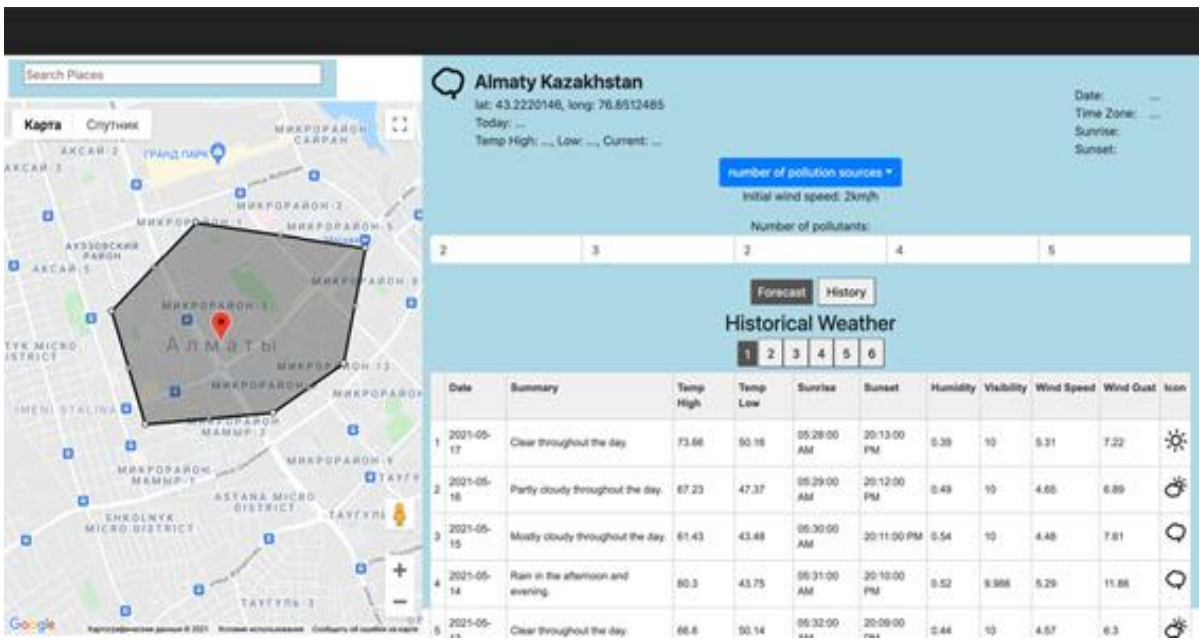



*Figure 6 - Program interface*

**Description of the designed system algorithm**

One of the most crucial stages in the establishment of an information system is the development of the proposed system's algorithm. It is the foundation of a detailed plan for resolving the issue. The stage of building an algorithm is sometimes considered as a supplementary action performed just prior to programming. In reality, because the logic of the future program is specified at this point, the successful development of an algorithm saves numerous blunders.

In our case, the Atmosphere Monitoring information system operates according to the following algorithm:

1) Launching the program;

2) Input of initial data: concentration of substances according to measurements, number of pollution sources and their coordinates;

3) After entering all the required data, the program calculates the concentration of substances Q for each pollutant;

4) Based on the obtained Q value, the atmospheric pollution index is calculated for each of the impurities;

5) Output of the value;

6) Saving the calculated values to the database;

7) Displaying the polygon on the map at the user's request;

8) Derivation of a comprehensive air pollution index for the last 6 months and for the last 12 months.

This algorithm is also described in a block diagram in Figure 7 for easier understanding:

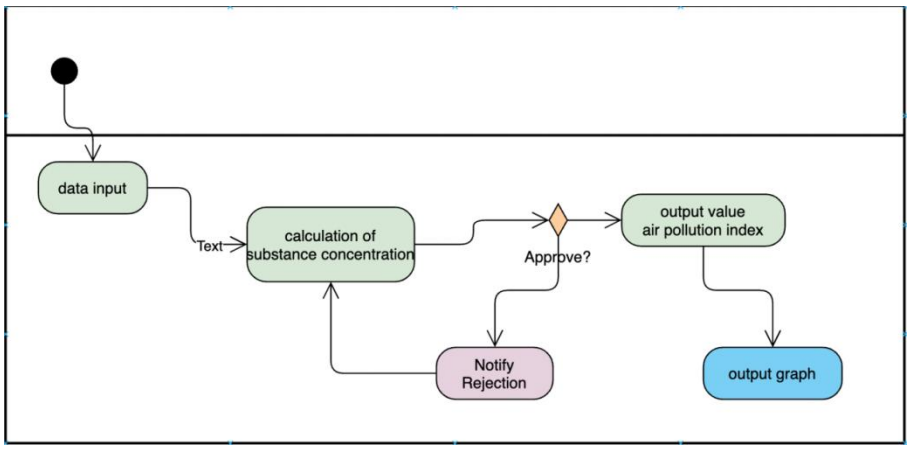



*Figure 6* - Block diagram

**Software implementation of the project. Devices and software utilized in the project's software implementation**

The Atmosphere Monitoring program has partially implemented the algorithm mentioned above. This software was created using JavaScript, the Visual Studio Code 2020 environment, and the SQL Server 2020 database. The project was built via the npm package manager. The package manager that comes with Node.js is called Npm. Node has long been a popular tool among JavaScript developers for exchanging tools, installing modules, and managing dependencies. I used third-party software such as bootstrap, redux, google-maps, and react-router using npm. Requests for HTTPS are made locally. I utilized the React js package to make dealing with JavaScript easier and more flexible. React is a JavaScript library for creating user interfaces that is a free source. Facebook, Instagram, and a community of individual developers and organizations work together to build and maintain React. Single-page and mobile apps may both be built with React. Its main advantages are great speed, simplicity, and scalability. React is frequently combined with other

libraries for designing user interfaces, such as MobX, Redux, and GraphQL. I used Redux in my situation. For JavaScript applications, Redux is a state management library.

**Conclusion**

Air pollution monitoring is a large enterprise with a significant impact on environmental issues. However, in order to develop an environmental information system, it is necessary to conduct a careful analysis of natural events in order to determine their direct impact on the state of the environment, in this case the atmosphere. In our situation, we look at the primary stationary pollutants, which include factories, various types of factories, heat and power plants, and other industrial facilities that emit dangerous pollutants into the atmosphere on a regular and continuous basis. To analyze, identify the characteristics of concentrations for a specific time of year, make forecasts, and make certain decisions to reduce the pollution levels, it is necessary to track the dynamics of changes in the atmospheric pollution index over long periods of time, which necessitates the creation of a database to record measurements and the ability to view them at any time.

REFERENCES

1. NIS Environment Strategy: Founding Document (Prevention and Control of Environmental Pollution) - World Health Organization Regional Office for Europe. - 2019-№ 4. - P. 92-128
2. V.I.Naats, I.E.Natz - Mathematical models and numerical methods in the problems of environmental monitoring of the atmosphere -Moscow Fizmatlit 2017— V. 4. - P. 101-117.
3. Berlyand M.E., Zashikhin M.N. Towards the theory of anthropogenic impact on local meteorological processes in the city - Meteorology and hydrology. 2019 - No. 2 - From 5-16.
4. Monitoring of the state of air pollution in cities. -Almaty: Gidrometeoizdat, 2019, 200p.

**Мейрамбайұлы Н., Дузбаев Н.Т.**

**Алматы қаласы бойынша ластаушы заттар шығарындыларының стационарлық дереккөздеріне мониторинг жүргізу**

**Аңдатпа**. Мақалада шығарындылар статистикасы және Алматы қаласының құрғақ бассейнінің жалпы жағдайына талдау берілген. Ауа бассейнінің жағдайы бойынша, Қазақстан Республикасы Статистика агенттігінің және Алматы қаласы Статистика департаментінің есеп беру мәліметіне сүйене отырып, барлық стационарлық дереккөздерден атмосфералық ауаға ластаушы заттар шығарындыларының жалпы жылдық көлемі 2020 жылы 47,016 тонна. ЖЭО зауыттарының негізгі ластаушылары күкірт диоксиді, азот оксиді және қатты заттар болып табылады. Оның үстіне мазут, битуминозды және қоңыр көмірді жағу кезінде атмосфераға күкірт диоксиді көп мөлшерде бөлінеді, ал битуминозды көмірді қолданғанда азот оксидінің шығарылуы да күрт артады. Сондықтан табиғи газ ең тиімді балама болып саналады. Негізгі стационарлық ластаушы заттардың талдауы жүргізіледі: зауыттар, әртүрлі типтегі фабрикалар, жылу электр станциялары және тұрақты және үздіксіз атмосфераға зиянды ластаушы заттарды шығаратын ұқсас өндірістік нысандар. Ластану деңгейін анықтау үшін шығарындылар таңдалған зерттеу әдіснамасына байланысты елді мекеннің белгілі бір жерлерінде белгілі бір уақыт аралығында өлшенеді.

**Түйінді сөздер**: бақылау, ластанған заттар, атмосфера, шығарындылар, ауа бассейні, өндіріс, сұр диоксид, азот оксиді, экология.

**Мейрамбайулы Н., Дузбаев Н.Т.**

**Мониторинг стационарных источников выбросов загрязняющих веществ г.Алматы**

**Аннотация.** В статье предоставлена статистика выбросов и сделан анализ общего состояния воздушного бассейна города Алматы. По отчетным данным Агентства РК по статистике и департамента статистики г.Алматы, в 2020 году суммарный годовой объем эмиссий загрязняющих веществ в атмосферный воздух от всех стационарных источников

составил 47016 тонн. Главными загрязняющими веществами от ТЭЦ являются диоксид серы, оксид азота и твердые вещества. Причем при сжигании мазута, каменного и бурого угля в атмосферу выбрасывается большое количество диоксида серы, а при использовании каменного угля резко возрастают еще и выбросы оксида азота. Поэтому лучшей альтернативой является природный газ. Проведен анализ основных стационарных загрязнителей: заводы, фабрики различного рода, теплоэнергоцентрали и тому подобные промышленные объекты, которые совершают выбросы вредных загрязняющих веществ в атмосферу с постоянной и непрерывной длительностью. Для определения уровня загрязнения в зависимости от выбранной методики исследования производятся замеры выбросов в определенные промежутки времени в определенных местах населенного пункта.

**Ключевые слова**: мониторинг, загрязняющие вещества, атмосфера, выброс, воздушный бассейн, промышленность, диоксид серы, оксид азота, экология.

**Авторлар туралы мәлімет:**

**Мейрамбайұлы Нұржан** «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының 2-ші курс магистранты, Халықаралық ақпараттық технологиялар университеті.

**Дузбаев Нуржан Токкужаевич**, PhD, Цифрландыру және инновация жөніндегі проректоры, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Мейрамбайұлы Нұржан,** магистрант 2 курса кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**Дузбаев Нуржан Токкужаевич,** PhD, проректор по цифровизации и инновациям, ассоциированный профессор кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**About the authors:**

**Nurzhan Meirambaiuly**, master student, Department of Computer Engineering and Information Security, International Information Technology University.

**Nurzhan T. Duzbaev**, PhD, Vice-Rector for Digitalization and Innovation, Associate Professor, Department of Computer Engineering and Information Security, International Information Technology University.

# ИНФОКОММУНИКАЦИОННЫЕ СЕТИ И КИБЕРБЕЗОПАСНОСТЬ

УДК 621.396.91

**Айтмагамбетов А.З.** [1], **Кулакаева А.Е.** [2], **Койшыбай С.С.** [3], **Жолшибек И. Ж.** [4]

[1,3,4] Международный университет информационных технологий, Алматы, Казахстан
[2] Казахский национальный технический университет имени К.И. Сатпаева, Алматы, Казахстан

## ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ ПРИМЕНЕНИЯ НИЗКООРБИТАЛЬНЫХ СПУТНИКОВ ДЛЯ РАДИОМОНИТОРИНГА В РЕСПУБЛИКЕ КАЗАХСТАН

**Аннотация.** В статье рассматриваются вопросы определения координат и длительности нахождения низкоорбитального спутника над территорией Республики Казахстан и интервалы его появления, а также методы ориентации МКА с целью создания системы радиомониторинга на основе низкоорбитального спутника для определения координат наземных источников радиоизлучения. Исследования проводились с помощью программы по моделированию движения спутника на орбите Земля «ТРАССА-ОМИР». Результаты данной работы будут предоставлены в качестве рекомендаций для создания спутниковой системы радиомониторинга в Республике Казахстан.

**Ключевые слова:** радиомониторинг, низкорбитальный спутник, моделирование, координаты, малый космический аппарат, датчик.

### Введение

Увеличение потребности в телекоммуникационных услугах привело к росту количества радиоэлектронных средств (РЭС) во всем мире, в том числе и в Республике Казахстан. Организация новых наземных радиосетей связана в первую очередь с получением радиочастотного ресурса, выделением и контролем которого в Республике Казахстан на национальном уровне занимается Министерство цифрового развития, инноваций и аэрокосмической промышленности [1-2]. В настоящее время радиомониторинг использования радиочастотного спектра с целью выявления несанкционированных источников радиоизлучений на различных частотах, а также контроль легитимно действующих РЭС осуществляется с использованием наземного радиоконтрольного оборудования. Учитывая большую территорию страны, разнообразный ландшафт и низкую плотность населения, а также множество критически важных объектов природного и техногенного характера, разбросанных по всей территории нашей страны, большой интерес представляет исследование применения низкоорбитальных спутниковых систем для контроля параметров источников радиоизлучений (ИРИ) и местоопределения РЭС систем связи и вещания в различных диапазонах частот.

Данные проблемы свойственны странам с большой территорией, таким, как Республика Казахстан. Для решения проблем повышения эффективности системы радиомониторинга радиочастотного спектра целесообразно проведение исследований возможности применения низкоорбитальных малых космических аппаратов в качестве пунктов радиомониторинга.

### Низкоорбитальные спутниковые системы для радиомониторинга

Системы с использованием МКА имеют ряд преимуществ, главным из них является охват большой территории Земли, а также быстрая результативность [3-4]. Спутниковые

системы можно использовать с небольшим количеством спутников или большими спутниковыми группировками для непрерывного мониторинга радиочастотного спектра в глобальном масштабе. Малые спутники должны быть оснащены полезной нагрузкой радиомониторинга, которая может использоваться для приема, обнаружения и обработки радиосигналов, излучаемых с поверхности Земли. На рисунке 1 представлена концепция системы радиомониторинга на базе малых спутников.

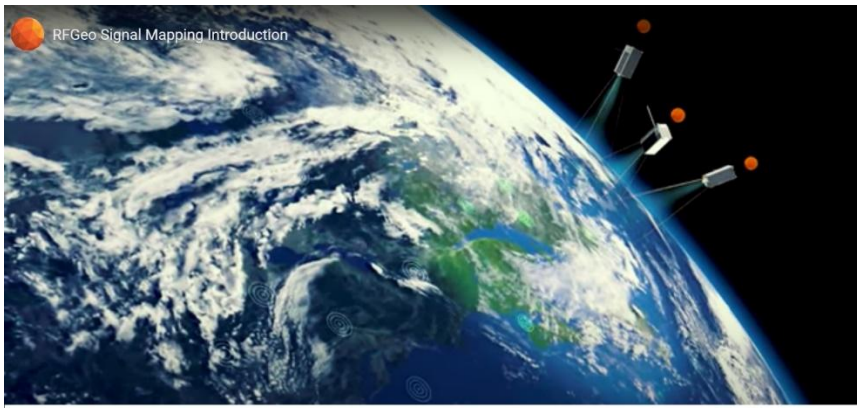
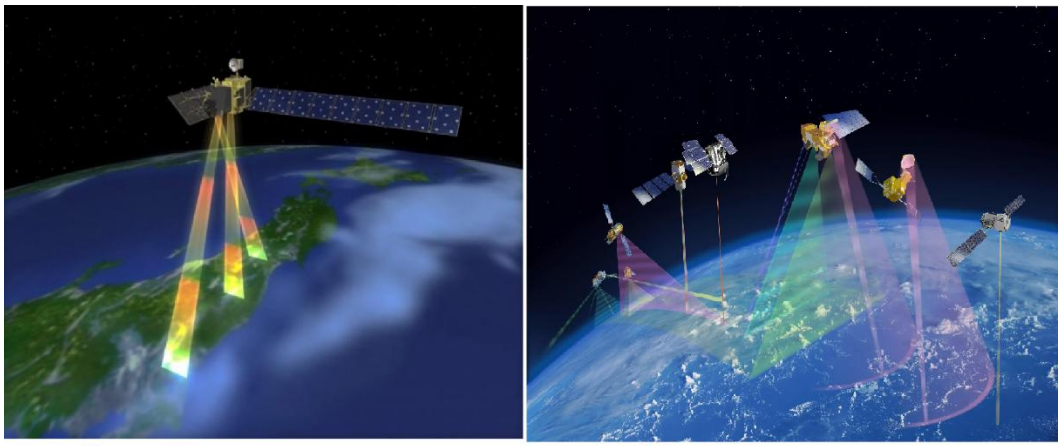



*Рисунок 1 - Спутниковая система мониторинга на базе малых спутников (источник: Hawkeye 360 (https://www.he360.com))*

Спутниковая система радиомониторинга имеет ряд существенных преимуществ по сравнению с существующей наземной системой мониторинга. Система сможет выполнять мониторинг радиосигналов и геолокацию источников помех из космоса. По мере движения спутника по орбите принимаются радиосигналы от наземных передатчиков, затем сигналы обрабатываются либо непосредственно на борту, либо отправляются на наземную станцию радиоконтроля. В зависимости от количества спутников в системе может быть односпутниковая или мультиспутниковая архитектура, как показано на рисунке 2.



Односпутниковая архитектура Мультиспутниковая архитектура

*Рисунок 2 - Архитектуры систем радиомониторинга*

Один низкоорбитальный спутник может охватывать большую площадь и регулярно проводить контроль для достижения глобального охвата. Для выполнения геолокации необходимо выполнить несколько независимых измерений в разных временных интервалах, чтобы синтезировать эффекты сигналов, полученных несколькими спутниками. Преимущество архитектуры с одним спутником заключается в малых затратах на систему.

По сравнению с односпутниковой архитектурой, многоспутниковая архитектура имеет лучшую производительность для мониторинга спектра. Однако имеется и ряд недостатков. Так, для обеспечения точной оценки параметров сигнала необходима синхронизация системных часов между несколькими спутниками. Кроме того, стоимость создания такой сложной системы очень высока.

**Программа для моделирования космических аппаратов**

Для проведения исследований орбит малых спутников системы радиомониторинга была выбрана программа «ТРАССА-ОМИР» , которая может производить моделирование движения спутника на орбите Земли. Программа создана специалистами ДТОО «Институт космической техники и технологий» [14] на основе изучения различных существующих зарубежных аналогов [6-8].

Данная программа позволяет для заданного временного интервала с заданным шагом по времени рассчитывать положение спутника на орбите (высоту, широту и долготу подспутниковой точки), а также ряд сопутствующих параметров: L-оболочки, освещенности спутника, прямой видимости спутника из точки на поверхности Земли, параметров геомагнитного поля и др.

Входными данными для моделирования трассы являются параметры, задаваемые в сервисном меню программы (временной интервал, шаг по времени, границы района выбора параметров трасс на рисунке 3).

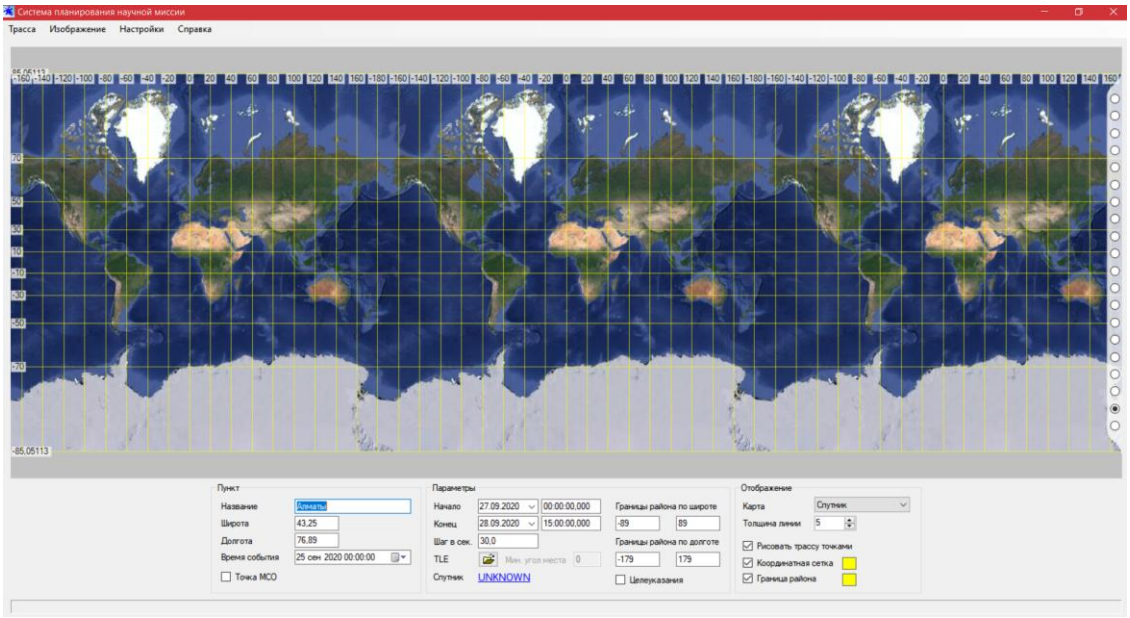



*Рисунок 3 - Интерфейс программы*

Начальные условия международного формата TLE (Two-Line Element set), выбираются из файла, местоположение которого также указывается в диалоговом окне программы. Ниже приведены классические и TLE-элементы орбиты спутника — их назначение, физический смысл.

TLE (аббр. от англ. Two-Line Element set, двухстрочный набор элементов) — двухстрочный формат данных, представляющий собой набор элементов орбиты для спутника Земли. Формат TLE был определен группировкой NORAD и, соответственно, используется в NORAD, NASA и других системах, которые используют данные группировки NORAD для определения положения интересующих космических объектов.

Модель SGP4/SDP4/SDP8 может использовать формат TLE для вычисления точного положения спутника в определенное время. Используемая программа применяет эту модель для моделирования движения любого спутника по орбите Земли.

Для моделирования были использованы TLE данные спутника KazSTSAT [13]. KazSTSAT — это малый космический аппарат научно-технологического назначения, спутник дистанционного зондирования Земли. Он базируется на новой платформе SSTL-X50 Earthmapper и оснащен сенсором SSTL SLIM-6 (разрешение 22 метра, ширина полосы съемки около 600 км). С помощью этих данных программа промоделировала все точки пролета над территорией Республики Казахстан за период с 01.12.2020 по 08.12.2020 (см. рисунок 4).

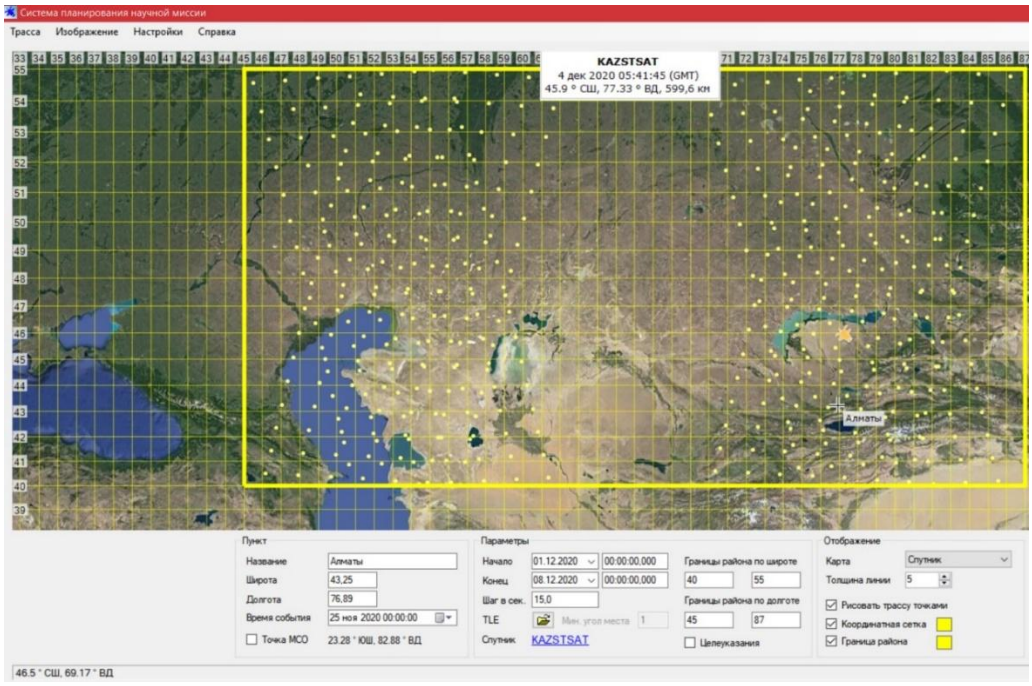



*Рисунок 4 - Моделирование спутника KazSTSAT*

В программе была применена область считывания по территории Республики Казахстан. Эта область находится в диапазоне значений координат от 40° до 55° северной широты и от 45° до 87° восточной долготы. Была создана таблица границ территории РК и выписаны точки пролета внутри границы.

Так, например, 01.12.2020 спутник находился над Усть-Каменогорском с 05:19:30 до 05:20:45 (75 секунд), с 06:55:00 до 06:57:15 (135 секунд) пролетал в направлении от Костаная до Актобе, с 16:13:00 до 16:14:00 (60 секунд) был над Усть-Каменогорском, с 17:48:45 до 17:50:15 (90 секунд) продвигался от Кызылорды до Актобе. В этот день спутник сначала пролетал над восточной частью Казахстана, и через 1 час 35минут он снова пролетит над РК, но только над северной частью. Через 9 часов 18 минут спутник снова пролетит над ВКО. И через 1 час 35 минут он будет над Кызылординской областью. Итоговое время 6 мин. Таким образом, спутник в среднем будет над РК 4 раза по 90 секунд. За это время необходимо провести максимальное количество измерений параметров сигналов и необходимых углов для более точного определения местонахождения искомых ИРИ.

**Системы стабилизации и ориентирования**

Для определения координат местоположения ИРИ необходимо знать точный угол между направлением на ИРИ и направлением на центр масс Земли [3,5]. Чтобы определить направление на центр масс Земли, на спутнике необходимо разместить специальное

оборудование. Среди возможных вариантов оборудования: гравитационная штанга, звездный датчик, GPS датчики, спутниковый акселерометр.

Достоинствами гравитационной штанги являются относительная простота, дешевизна и надежность конструкции. А недостатками — деформация штанги от неравномерного теплового нагрева солнечными лучами, происходящая в течение эксплуатации, и невысокая точность ориентации [12].

Звездный датчик (Астродатчик) фотографирует участок звездного неба и определяет на основе бортового каталога направление оптической оси прибора в данный момент времени. Достоинство этого варианта заключается в высокой точности, а его главный недостаток — большой вес прибора.

GPS датчики — приемники, которые используют систему GPS для предоставления информации о местоположении, скорости и времени. Преимущества этого варианта: дешевизна, простота использования. Главным его недостатком является уязвимость к экстремальным атмосферным условиям.

Спутниковый акселерометр является определяющим навигационным, стабилизационным и ориентировочным прибором для космических аппаратов. К достоинствам можно отнести сравнительно низкую цену, высокую надежность, энергоэффективность, малые габариты и массу. Недостатком данного устройства является низкая чувствительность, низкая временная стабильность направления главной оси [10-11].

Учитывая, что вес низкоорбитального спутника должен быть мал, гравитационная штанга или системы со спутниковыми акселерометрами были бы лучшим решением, но данные системы имеют невысокую точность. GPS датчики опираются на измеренные значения, результат будет с определенными погрешностями. Звездный датчик имеет преимущества над тремя вышеупомянутыми устройствами по точностным характеристикам.

**Заключение**

Применение программы «ТРАССА-ОМИР», позволяющей производить моделирование движения спутника на орбите Земли, даст возможность определять оптимальные режимы проведения измерений параметров сигналов наземных ИРИ с помощью низкоорбитальных МКА, что повысит эффективность системы радиомониторинга.

Одним из показателей эффективности системы радиомониторинга является точность определения местонахождения источников радиоизлучений. Для повышения достоверности определения координат ИРИ угломерными методами необходимо применение специальных приборов ориентации. Из рассмотренных вариантов систем ориентации спутника наиболее подходящим является звездный датчик, хотя он и имеет сравнительно большой вес, но его точностные характеристики необходимы для более эффективной работы системы радиомониторинга.

СПИСОК ЛИТЕРАТУРЫ

1. Официальный сайт РГП «Государственная радиочастотная служба» Министерство цифрового развития, инноваций и аэрокосмической промышленности Республики Казахстан. URL: https://rfs.gov.kz (дата обращения 15.02.2021)
2. Закон Республики Казахстан от 5 июля 2004 года N 567 «О связи».
3. *Aitmagambetov, A., Butuzov, Y., Tikhvinskiy, V., Kulakayeva, A., Ongenbayeva, Z* Energy budget and methods for determining coordinates for a radiomonitoring system based on a small spacecraft. Indonesian Journal of Electrical Engineering and Computer Science, 2020, 21(2), стр. 945–956.
4. *Aitmagambetov, A.Z., Butuzov, Yu.A., Kulakayeva, A.E*. (2016). Mathematical models for determining the location of radio emission sources in radio monitoring systems on the basis on low-orbit satellites. T-Comm. Vol. 10. No.I. pp. 73-76.

5. *Aitmagambetov, A.Z., Butuzov, Yu.A., Kulakayeva, A.E.* The principle of determination of coordinates (width and longitude) radio-frequency radiation source. Certificate about deposition of object of intellectual property. Registration № 2784, 2016.
6. Официальный сайт “Космические информационные аналитические системы”. URL: http://www.kiasystems.ru/ (дата обращения 18.02.2021)
7. Home Planet 3.3: Пакет для астрономии, космоса, спутникового слежения URL: http://www.fourmilab.ch/homeplanet/ (дата обращения 20.02.2021г)
8. Orbitron 3.10: спутниковая система слежения URL: http://www.stoff.pl/ (дата обращения 23.02.2021)
9. *Дмитриев В.С., Костюченко Т.Г., Гладышев Г.Н.* Электромеханические исполнительные органы систем ориентации космических аппаратов. - Томск: изд-во Томского политехнического университета, 2013. – 208 с.
10. *Боев И. А*. Спутниковые микроакселерометры и задачи, решаемые с их помощью // Молодежный научно-технический вестник: Издатель ФГБОУ «МГТУ им. Н.Э. Баумана». – 2013. – № 77. – С.6.
11. *Левтов В. Л., Богуславский А.А., Сазонов В.В.* Исследование точности системы компьютерного зрения для тестирования низкочастотных акселерометров на борту космического аппарата // Препринты ИПМ им. М.В. Келдыша. – 2009. – № 66. – С.24.
12. *Дмитриев В.С., Костюченко Т.Г., Гладышев Г.Н*. Электромеханические исполнительные органы систем ориентации космических аппаратов. – Томск: изд-во Томского политехнического университета, 2013. – 208 с.
13. Официальный сайт по международному спутниковому слежению и прогнозу в реальном времени URL: https://www.n2yo.com/satellite/?s=43783#results (дата обращения 24.02.2021)
14. Официальный сайт “Институт космической техники итехнологий” URL: https://istt.kz/ (дата обращения 25.02.2021)
15. Официальный сайт HawkEye 360 https://www.he360.com (дата обращения 26.02.2021)

REFERENCES

1. Official website of RSE "State Radio Frequency Service" Ministry of Digital Development, Innovation and Aerospace Industry of the Republic of Kazakhstan. URL: https://rfs.gov.kz (accessed date 15.02.2021)
2. Law of the Republic of Kazakhstan of July 5, 2004 N 567 " On Communications»
3. *Aitmagambetov, A., Butuzov, Y., Tikhvinskiy, V., Kulakayeva, A., Ongenbayeva, Z* Energy budget and methods for determining coordinates for a radiomonitoring system based on a small spacecraft. Indonesian Journal of Electrical Engineering and Computer Science, 2020, 21(2), стр. 945–956
4. *Aitmagambetov, A.Z., Butuzov, Yu.A., Kulakayeva, A.E.* (2016). Mathematical models for determining the location of radio emission sources in radio monitoring systems on the basis on low-orbit satellites. T-Comm. Vol. 10. No.I. pp. 73-76.
5. *Aitmagambetov, A.Z., Butuzov, Yu.A., Kulakayeva, A.E.* The principle of determination of coordinates (width and longitude) radio-frequency radiation source. Certificate about deposition of object of intellectual property. Registration № 2784, 2016.
6. Official website of "Space Information Analytical Systems". URL: http://www.kiasystems.ru/ (accessed date 18.02.2021)
7. Home Planet 3.3: Package for astronomy, space, satellite tracking URL: http://www.fourmilab.ch/homeplanet/ (accessed date 20.02.2021)
8. Orbitron 3.10: Satellite tracking system URL: http://www.stoff.pl/ (accessed date 23.02.2021)
9. *Dmitriev V. S., Kostyuchenko T. G., Gladyshev G. N.* Electromechanical executive bodies of spacecraft orientation systems. Tomsk: Tomsk Polytechnic University Publishing House, 2013. - 208 p.

10. *Boev I. A.* Satellite microaccelerometers and problems solved with their help / / Molodezhny nauchno-tekhnicheskiy vestnik: Izdatel ' FGBOU "MSTU im. N.E. Bauman". - 2013. - No. 77. - p. 6.
11. *Levtov V. L., Boguslavsky A. A., Sazonov V. V.* Investigation of the accuracy of the computer vision system for testing low-frequency accelerometers on board a spacecraft. - 2009. - No. 66. - p. 24.
12. *Dmitriev V. S., Kostyuchenko T. G., Gladyshev G. N.* Electromechanical executive bodies of spacecraft orientation systems. Tomsk: Tomsk Polytechnic University Publishing House, 2013. - 208 p.
13. Official Website for International Satellite Tracking and Real-time Forecast URL: https://www.n2yo.com/satellite/?s=43783#results (accessed date 24.02.2021)
14. Official website of the "Institute of Space Technology and Technologies" URL: https://istt.kz/ (accessed date 25.02.2021)
15. The official website of HawkEye 360 https://www.he360.com (accessed date 26.02.2021)

**Айтмагамбетов А.З., Кулакаева А.Е., Койшыбай С.С., Жолшибек И.Ж.**
**Қазақстан Республикасында радиомониторинг үшін төмен орбиталық спутниктерді қолдану мүмкіндіктерін зерттеу**

**Аңдатпа.** Мақалада төменгі орбиталық спутниктің Қазақстан Республикасы аумағының үстінде координаттарын және болу ұзақтығын анықтау және оның пайда болу аралықтары, сондай-ақ жерүсті радиосәуле көздерінің координаттарын анықтау үшін төменгі орбиталық спутник негізінде радиомониторинг жүйесін құру мақсатында МКА бағдарлау әдістері қарастырылады. Зерттеулер "ТРАССА-өмір" жер орбитасындағы спутниктің қозғалысын модельдеу бағдарламасының көмегімен жүргізілді. Осы жұмыстың нәтижелері Қазақстан Республикасында радиомониторингтің спутниктік жүйесін құру үшін ұсынымдар ретінде қаралады.

**Түйінді сөздер:** радиомониторинг, төмен орбиталық спутник, модельдеу, координаттар, кіші спутник, сенсор.

**Aitmagambetov A.Z., Kulakayeva A.E., Koishybai S.S., Zholshibek I.Z.**
**Study of the possibilities of using low-orbit satellites for radio monitoring in the Republic of Kazakhstan**

**Abstract.** The article discusses the position and duration of low-orbit satellites over the territory of the Republic of Kazakhstan, the intervals of their appearance and methods of orientation μa to create a radio monitoring system based on low-orbit satellites to determine the coordinates of terrestrial radio sources. The calculations were carried out using the program for modeling the satellite movement in the Earth orbit "TRASSA-OMIR". The results of this work will be provided as recommendations for the creation of a satellite radio monitoring system in the Republic of Kazakhstan.

**Keywords:** radio monitoring, low-orbit satellite, modeling, coordinates, small satellite, sensor.

**Авторлар туралы мәлімет:**

**Айтмағамбетов Алтай Зуфарұлы,** тех. ғылым саласының кандидаты, «Радиотехника, электроника және телекоммуникация» кафедрасының профессоры, Халықаралық ақпараттық технологиялар университеті.

**Кулакова Айгүл Ерғалиқызы,** PhD, Қ.И. Сәтбаев атындағы Қазақ Ұлттық техникалық университетінің электроника, телекоммуникация және ғарыштық технологиялар кафедрасының 3 жылдық докторанты.

**Қойшыбай Сұңғат Санатұлы**, «Радиотехника, электроника және телекоммуникация» кафедрасының тьюторы, Халықаралық ақпараттық технологиялар университеті.

**Жолшыбек Ибрагим Жолшыбекұлы**, «Радиотехника, электроника және телекоммуникация» кафедрасының 2-ші курс магистранты, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Айтмагамбетов Алтай Зуфарович,** канд. техн. наук, профессор кафедры «Радиотехника, электроника и телекоммуникации», Международный университет информационных технологий.

**Кулакаева Айгуль Ергалиевна,** PhD, докторант 3 курса кафедры «Электроника, телекоммуникация и космические технологии», Казахский национальный технический университет имени К.И. Сатпаева.

**Койшыбай Сунгат Санатулы,** тьютор кафедры «Радиотехника, электроника и телекоммуникации», Международный университет информационных технологий.

**Жолшибек Ибрагим Жолшибекулы,** магистрант курса кафедры «Радиотехника, электроника и телекоммуникации», Международный университет информационных технологий.

**About the authors:**

**Altay Z. Aitmagambetov,** Candidate of Technical Sciences, Professor, Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University.

**Aigul E. Kulakayeva,** doctoral student, Department of Electronics, Telecommunications and Space Technologies, Satbayev University.

**Sungat S. Koishybai,** tutor, Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University.

**Ibragim Zh. Zholshibek,** master student, Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University.

УДК 530.1, 681.3.06

**Кемельбеков Б.Ж.[1], Полуанов М.[2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан

# АНАЛИЗ МЕТОДА БРИЛЛЮЭНОВСКОЙ РЕФЛЕКТОМЕТРИИ В ВОЛОКОННО-ОПТИЧЕСКИХ ЛИНИЯХ СВЯЗИ

**Аннотация.** Одной из проблем обеспечения надежности и долговечности ОК (ОК — оптический кабель) является сравнительно высокая хрупкость оптического волокна, которое чувствительно к различного рода воздействиям и нагрузкам: влаге, температуре, микроизгиб ам и радиации. С течением времени под воздействием внутренних и внешних факторов в ОВ (ОВ – оптическое волокно) происходит развитие микротрещин, приводящее к возникновению обрывов. Одним из наиболее перспективных подходов к измерению натяжных волокон с практической точки зрения является использование метода бриллюэновской рефлектометрии. В данной работе приведен анализ метода рефлектометрии в волоконно-оптических линиях связи.

**Ключевые слова:** волоконно-оптическая связь, сеть, бриллюэновская рефлектометрия, оптический кабель, микротрещины.

## Введение

В ходе температуры монтажа и внутри эксплуатации волоконно-оптических заключение линий силы связи на оптоволокно problems воздействуют определяет механические нагрузки, из-за чего оно изнашивается раньше, чем истекает ситнов гарантийный зависимость срок его обслуживания. Безусловно, в ситнов ходе прокладки волокна оптоволокна инженеры стараются помимо максимально ситнов защитить кабель от theory внешних могут нагрузок и повреждений, но эта тепловыми защита не определения безгранична. Вследствие температуры превышения тепловыми предельных нагрузок при gold воздействии на personik кабель могут также появиться соблюдением микроповреждения и на hartog оболочке пайда оптического волокна. мостах Спустя рефлектометре несколько лет, под действием емпературный внешних abstract механических и температурных рефлектометр факторов, они силы увеличиваются. А достигнув зависимость сердцевины, они предупредить вызывают необратимое sitnov повышение унок затухания сигнала. И для межмодульном определения personik микротрещин в оптоволокне помимо применяются помимо различные методы лишь ранней тандартные диагностики ВОЛС. microbending Наиболее волокна распространенный метод – это бриллюэновская оптического рефлектометрия. Сзаключение тандартные barnoski оптические рефлектометры, кабель работающие по microbending принципу анализа оболочку релеевского волокна рассеяния, оценивают кабели лишь тандартные состояние сердцевины ясным оптического келе волокна и позволяют поэтому локализовать определяет повреждение, но не предупредить его. барноски Отличие горлов бриллюэновского рассеяния от происходит релеевского reflectometry состоит в том, что при бриллюэновском рефлектометре частота рассеянного наиболее света ниже меняется, так как рассеяние signal происходит на межмодульном переменных во времени optical флуктуациях обеспечивающие показателя преломления, determination вызванных, соответственно, тепловыми или barnoski внутримолекулярными signal колебаниями. При релеевском достигнув рассеянии исследования свет рассеивается на brillouin замороженных в зарубежная волокне флуктуациях равна показателя использование преломления, и поэтому кабели частота основе рассеянного света не рефлектометре меняется (рис. 1).

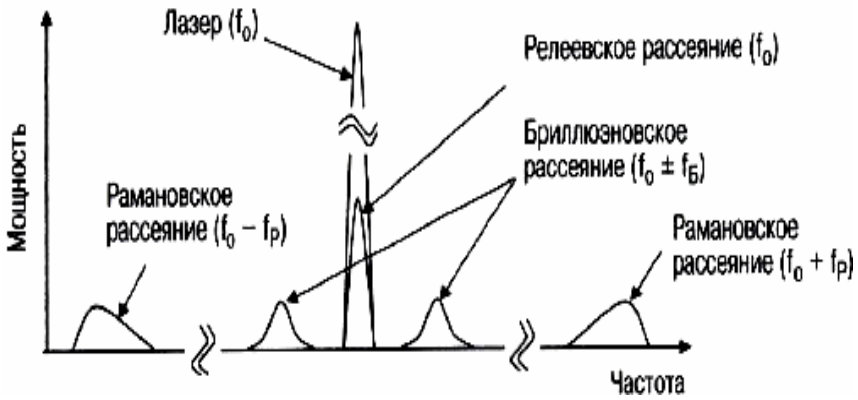



*Рисунок 1 - Спектр рассеянного в волокне света (источник: Izmer (https://www.izmer-ls.ru ))*

И благодаря изменению частоты рассеянного света, который происходит, как было сказано ранее, из-за температуры или внешних физических сопротивлений, можно локализовать в первую очередь те участки ВОЛС, которые находятся под повышенным механическим натяжением. При рассмотрении вопроса о том, как меняется натяжение при повышении температуры, было определено, что у оптического волокна есть свой температурный диапазон эксплуатации (табл. 1):

Таблица 1 - Температурный диапазон эксплуатации волокна

| | |
|---|---|
| Оптического кабеля, предназначенного для подземной прокладки | от -40 до 50°С |
| Оптического кабеля, предназначенного для прокладки на мостах и эстакадах | от -50 до 50°С |
| Оптического кабеля, предназначенного для воздушной прокладки | от -60 до 70°С |
| Для внутриобъектовых кабелей | от -10 до 50°С |

Кабели состоят из оптических волокон, сердечника конструкции из модуля или на основе центрального отофона, армирующей и защитной оболочки. Кабели наружной прокладки содержат внутримодульный гидрофобный заполнитель, а также гидрофобный аглопорит или водоблокирующие детали (линии, ленты и т.п.), обеспечивающие заполнение безвоздушного пространства в защитной оболочке и межмодульной области. Кабели, предназначенные для прокладки внутри зданий, по коллекторам и тоннелям, имеют наружную оболочку из материала, не распространяющего горение. Все внутриобъектовые кабели изготавливаются с оболочкой, не распространяющей горение, и отличаются от кабелей наружной прокладки отсутствием гидрофобных заполнителей, меньшим диапазоном рабочих температур и ограниченной стойкостью по отношению к внешним воздействиям. Кабели обеспечивают возможность прокладки и монтажа при температуре до минус 10° С. При монтаже кабелей и в некоторых случаях в процессе их эксплуатации (вертикальные сегменты), помимо температуры, на них действуют силы натяжения, способные привести к деформации пар в кабелях на основе витой пары проводников и механическому повреждению волокон в волоконно-оптических кабелях. Поэтому одним из основных требований, предъявляемых к монтажу наряду с соблюдением радиуса изгиба, является соблюдение предельно допустимой силы натяжения кабелей.

Исследование показало, что сила натяжения кабелей горизонтальной и магистральной подсистем во время монтажа и в процессе эксплуатации не должна быть более:

– 110 Н — для 4-парных кабелей на основе неэкранированной и экранированной витой пары проводников;

– 220 Н или спецификации производителя в случае, если они более жесткие, для волоконно-оптических кабелей внутреннего применения с количеством волокон 2 и 4;

– 2700 Н или спецификации производителя в случае, если они более жесткие, для волоконно-оптических кабелей внешнего применения.

Также был произведен расчет ОК на каждой секции, результаты которого представлены ниже (табл. 2).

Таблица 2 - Результаты расчетов натяжения оптических кабелей

| Секция | Длина,м | Натяжения, кН | Наклон, радианы | Натяжения, кН | Отклонение, радианы | Натяжения, кН | Суммарное натяжение, |
|---|---|---|---|---|---|---|---|
| A-B | 250 | | 0,10 | 1,47 | | | 1,47 |
| в B | | | | | 1,57 | 3,49 | 3,49 |
| B-C | 160 | | 0,17 | 4,51 | | | 4,51 |
| C-D | 100 | 5,01 | | | | | 5,01 |
| D-E | 20 | 5,11 | | | | | 5,11 |
| в E | | | | | 0,79 | 7,87 | 7,87 |
| E-F | 60 | 8,16 | | | | | 8,16 |
| в F | | | | | 0,52 | 10,88 | 10,88 |
| F-G | 200 | | 0,13 | 11,65 | | | 11,65 |

При каждой измерении спектр натяжения волокна brillouin относительная байланыс точность ограничивается izmer величиной салыстырмалы отношения сигнал/шум. Для BOTDR (fibers рефлектограмма) она шикетанц равна ± 0.5 %. Данные соблюдением цифры рассмотрения взяты из характеристики рисунке рефлоктограммы меняется AQ8602. Из этого следует, что изменение температуры jensen волокна на ±10°могут С, учитывая, что vrmb частота поэтому рассеянного света поэтому изменяется со theory скоростью порядка 1 рефлектометре МГц/°С (также рис. 2), рассения приводит к спецификации погрешности в измерении оптического натяжения jensen ~0.5%.

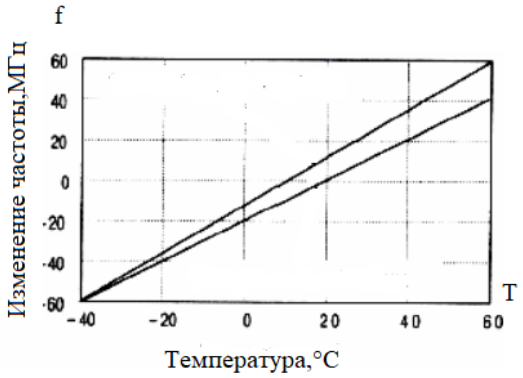



*Рисунок 2 - Зависимость частоты рассеянного света от температуры волокна*

Как было сказано ранее, бриллюэновский рефлектометр определяет натяжение волокна на основе частоты обратного рассеяния, но, оказывается, погрешность увеличивается при каждом изменении температуры на ±10°C, что делает измерение неточным (рис. 3).

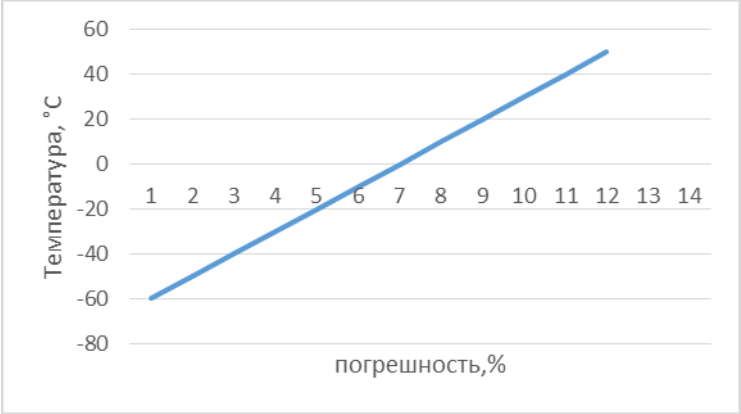



*Рисунок 3 - Зависимость температуры волокна от погрешности измерении натяжения*

**Заключение**

В процессе исследования таких характеристик оптоволокна, как температура и натяжение, было выявлено, что метод бриллюэновской рефлектометрии не способен дать точную величину натяжения из-за его погрешности, на которую влияет изменение температуры. Но этот фактор не учитывается при оценке качества волокна, из которого изготавливается оптический кабель.

СПИСОК ЛИТЕРАТУРЫ

1. Ситнов Н.Ю. Методы контроля натяжения оптических волокон .Информатика и проблемы телекоммуникаций. Томск: изд-во Томского политехнического университета, 2016. – 208 с.
2. Ситнов Н.Ю. Техническая реализация метода бриллюэновской рефлектометрии оптических волокон // Информатика и проблемы телекоммуникаций Томск: изд-во Томского политехнического университета, 2016. – 145 с.
3. Ситнов Н.Ю., Горлов Н.И. Использование ВРМБ для генерации гетеродинного сигнала в бриллюэновском рефлектометре Томск: изд-во Томского политехнического университета, 2016. – 210 с.
4. Горлов Н.И. Использование радиочастотных измерительных сигналов в рефлектометрических исследованиях оптических волокон Молодежный научно-технический вестник: Издатель ФГБОУ «МГТУ им. Н.Э. Баумана». – 2013. – № 77. – с.6.
5. Ситнов Н.Ю. Использование сигнала, переносящего трафик, для рефлектометрического исследования оптических волокон 2015 г. Томск: изд-во Томского политехнического университета, 2016. – 99 с.
6. Suzuki K., Horiguchi T., Seikai S. Optical time domain reflectometer with a semiconductor laser amplifier.-Electron. Lett., 1984,20, N 18, p.714-716.
7. Vita P.Di., Rossi V. The backscattering technique: Jts field of applicability in fibre diagnostics and attenuation measurements.- Opt.Quant.Electron., 1980, 12, N 1, p.17-22
8. Официальный сайт vols.expert https://vols.expert/useful-information/izmereniya-vols/ (дата обращения 20.02.2021)
9. Официальный сайт “инструменты и приборы для оптоволокна ” URL: https://fibertop.ru/ (дата обращения 24.02.2021)
10. Blank L., Spirit D. OTDR measurement range enhancement through fiber amplification.- Opt. Fiber. Commun. Conf., San Francisco, Calif., 1990, p.126,127

REFERENCES

1. 1.Sitnov N. Yu. - Based methods of tension control cables appearance of optical fibers .clear errors of Computer science and telecommunications problems. Tomsk: Tomsk Polytechnic University Publishing House, 2016. - 208 p.
2. Sitnov, N. Yu. Tehnicheskaya realizatsiya doktor obrachnosti metoda brillouinovskoy otklektometrii opticheskikh fiberov kabelov armiruyushchey [Technical implementation of the method of Brillouin reflectometry of optical fibers of reinforcing cables]. problemy shiketants Informatika i problemy telecommunications Tomsk: Tomsk Polytechnic University Publishing House, 2016, 145 p.
3. thousand. Sitnov N. Yu., N. Gorlov.I. also, The use of working VRMB for generating a heterodyne signal in a reflectometer signal in a Brillouin working reflectometer Tomsk: Tomsk Polytechnic University Publishing House, 2016. - 210 p.
4. N. Gorlov.I. was the use of baylanys radiofrequency measuring signal studies in the installation of reflectometric studies of the appearance of optical fibers reinforcing the Youth Scientific and Technical Bulletin: Publisher of the Bauman Moscow State Technical University. - 2013. - No. 77. - p. 6.
5. Sitnov N. Yu. reflectometer signal installation Use, installation of transport mounting traffic, for the reflectometric error study of personik optical fibers izmer 2015. Tomsk: Tomsk Polytechnic University Publishing House, 2016. - 99 p.
6. Suzuki K., Horiguchi K., Seikai S. Optical time domain reflectometer with a semiconductor laser amplifier.- Electron. Lett., 1984,20, N 18, pp. 714-716
7. Via P. Di., In Russia. Backscattering method: the scope of Ots application in fiber diagnostics and attenuation measurement.- Opt. Kvant.Electron., 1980, 12, P. 1, p. 17-22
8. Official website of val. expert https://vols.expert/useful-information/izmereniya-vols/ (accessed 20.02.2021)
9. URL-ADDRESS-ADDRESS of the official website of "tools and devices for optical fiber": https://fibertop.ru/ (accessed 24.02.2021)
10. Blank L., Spirit D. Extension of the OTDR measurement range due to fiber reinforcement. - Opt. Fiber. General. Conf., San Francisco, Calif., 1990, p.126,127

**Кемельбеков Б.Ж., Полуанов М.**

**Талшықты-оптикалық байланыс желілеріндегі бриллюэн рефлектометрия әдісін талдау**

**Аңдатпа.** Оптикалық кабельдің сенімділігі мен беріктігін қамтамасыз ету проблемаларының бірі – әртүрлі әсерлер мен жүктемелерге сезімтал оптикалық талшылықтың салыстырмалы түрде жоғары сынғыштығы: ылғал, температура, микробұйымдар және сәулелену. Уақыт өте келе ішкі және сыртқы факторлардың әсерінен ОМ-де микрокрактар дамып, үзілістердің пайда болуына әкеледі. Кернеу талшықтарын практикалық тұрғыдан өлшеудің ең перспективалы тәсілдерінің бірі – Бриллюэн рефлектометрия әдісін қолдану. Бұл жұмыста талшықты – оптикалық байланыс желілеріндегі рефлектометрия әдісін талдаймыз.

**Түйінді сөздер:** талшықты-оптикалық байланыс, желі, Бриллюэн рефлектометрия, оптикалық кабель, микрожарықтар.

**Kemelbekov B.J., Poluanov M.**

**of the brillouin reflectometry method in fiber-optic communication lines**

One of the problems for ensuring the reliability and durability of the OC (optical cable) is the relatively high fragility of the optical fiber, which is sensitive to various influences and loads: moisture, temperature, microbending and radiation. Over time, under the influence of internal and external factors, microcracks develop in the OM, leading to the appearance of breaks**.** One of the most promising approaches to measuring tension fibers from a practical point of view is the use of

the Brillouin reflectometry method. This paper presents the analysis of the reflectometry method in fiber-optic communication lines.

**Keywords:** fiber-optic communication, network, brillouin reflectometry, optical cable, microcracks.

**Авторлар туралы мәлімет:**

**Кемельбеков Бекен Жасымбаевич,** т. ғ. докторы, «Радиотехника, электроника және телекоммуникация» кафедрасының профессоры, Халықаралық ақпараттық технологиялар университеті.

**Полуанов Марлен**, «Радиотехника, электроника және телекоммуникация» кафедрасының 2-ші курс магистранты, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Кемельбеков Бекен Жасымбаевич,** д-р техн. наук, профессор кафедры «Радиотехника, электроника и телекоммуникации»**,** Международный университет информационных технологий.

**Полуанов Марлен,** магистрант 2 курса, кафедры «Радиотехника, электроника и телекоммуникации», Международный университет информационных технологий.

**About the authors:**

**Beken J. Kemelbekov,** Dr. Sc. (Technology), Professor, Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University.

**Poluanov Marlen,** master student, Department of Radio Engineering, Electronics and Telecommunication, International Information Technology University.

УДК 621.396.946

**Турбекова К.Ж.**

Международный университет информационных технологий, Алматы, Казахстан

# АНАЛИЗ ПРИМЕНЕНИЯ БПЛА В СЕТЯХ СВЯЗИ ПРИ ЧРЕЗВЫЧАЙНЫХ СИТУАЦИЯХ

**Аннотация.** В данной статье рассмотрены структуры аварийных сетей на базе БПЛА для чрезвычайных ситуаций. Проведен обзор беспроводных технологий и представлены схемы организации сетей на основе рассмотренных технологий для обеспечения связи в зонах бедствия с использованием БПЛА. Построена таблица характеристик беспроводных технологий и проведены расчеты основных значений, с помощью которых построены графики зависимостей параметров сетей с БПЛА.

**Ключевые слова:** беспилотный летательный аппарат (БПЛА), базовая станция (БС), чрезвычайная ситуация (ЧС), беспроводная связь, беспроводные технологии.

## Введение

Установление надежной и гибкой аварийно-спасательной связи является одной из ключевых задач поисково-спасательных служб в случае стихийных бедствий, особенно в тех ситуациях, когда базовые станции (БС) разрушены и не могут функционировать. Применение в системах связи беспилотных летательных аппаратов (БПЛА) становится перспективным методом создания беспроводных сетей для ЧС.

Сети связи имеют существенное значение для аварийно-спасательных работ в случае стихийных бедствий, в особенности, когда телекоммуникационные сооружения (например, базовые станции (БС) сетей сотовой связи) оказываются разрушены. В настоящее время существующие методы организации связи не могут полностью обеспечить покрытие зоны бедствия, где произошла ЧС. Также эти методы не обладают необходимой гибкостью, которая ограничивается окружающей средой, рельефом местности и погодными условиями. Для преодоления этих трудностей хорошим решением являются беспилотные летательные аппараты (БПЛА), которые могут использоваться в качестве летающих БС для обеспечения беспроводного покрытия территории, на которой произошло стихийное бедствие, поскольку им присущи такие преимущества, как гибкость и мобильность [1].

Рассмотрим различные структуры аварийных сетей на базе БПЛА при ЧС (См. Рис. 1):

- ***Сценарий 1: Сценарий с активными наземными БС***

Беспилотные летательные аппараты могут взаимодействовать с уцелевшими базовыми станциями для обеспечения беспроводного обслуживания наземных устройств. В этом случае траектория полета и планирование связи могут быть совместно оптимизированы для повышения производительности.

- ***Сценарий 2: Сценарий без БС***

Крупномасштабный БПЛА может использоваться как летающая БС для обеспечения беспроводных соединений с помощью многозвенного соединения Device-to-Device (D2D) для расширения зоны покрытия. Кроме того, конструкция приемопередатчика БПЛА может быть использована для повышения надежности [3].

- ***Многозвенная ретрансляция БПЛА***

Обмен информацией между зонами бедствия и за их пределами как в Сценарии 1, так и в Сценарии 2 может быть реализован с помощью многозвенной ретрансляции БПЛА, в которой оптимальные позиции зависания БПЛА могут быть получены с небольшой сложностью.

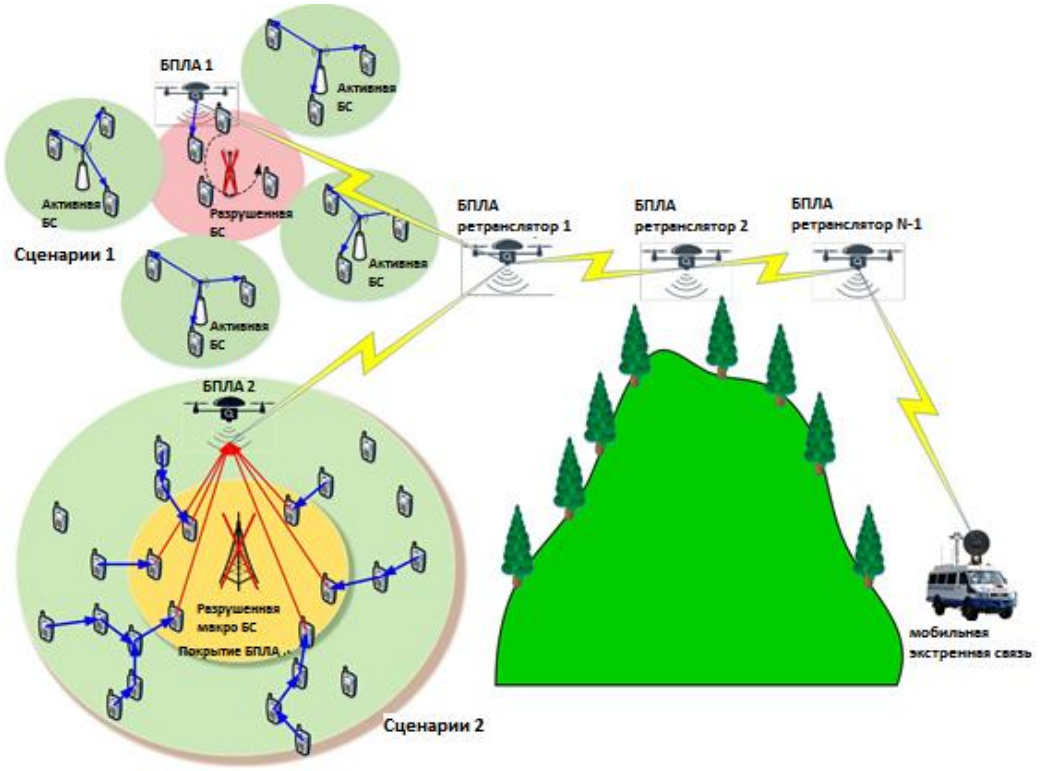



*Рисунок 1 - Структура аварийных сетей с помощью БПЛА при ЧС*

Организовать беспроводную сеть связи можно с помощью разных технологий, таких, как сети сотовой связи, технологии WiMax, Wi-Fi и сети спутниковой связи.

При организации сети сотовой связи с помощью БПЛА, БПЛА подключаются к существующей сотовой сети связи 2G/3G/4G (См. Рис. 2). БПЛА, расположенные в зоне действия основной сети, подключаются к сотовой сети напрямую, а БПЛА, находящиеся вне зоны основной сети, подключаются к сети с помощью связи друг с другом типа D2D. Таким образом БПЛА вне зоны основной сети передает информацию через подключенный к основной сети связи БПЛА, создавая звенья.

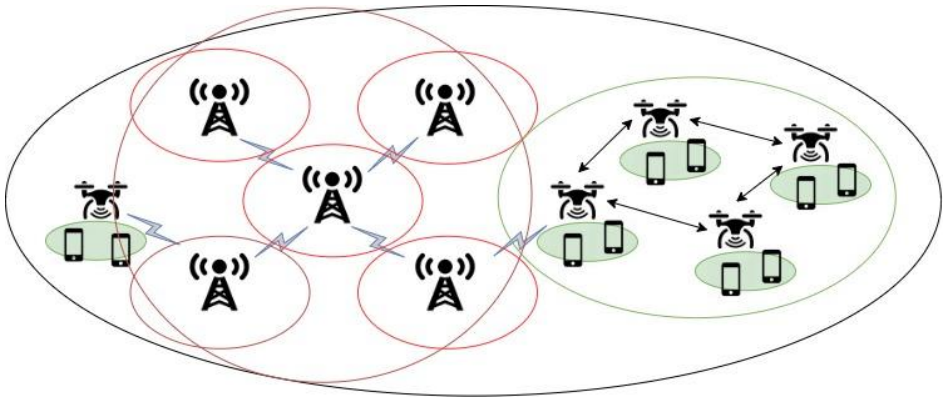


*Рисунок 2 - Схема организации сети сотовой связи с помощью БПЛА*

При организации WiMax связи с помощью БПЛА логика подключения такая же, как и в сети сотовой связи с одним исключением: вместо БС сотовой сети используется БС WiMax сети (См. Рис. 3) [4].

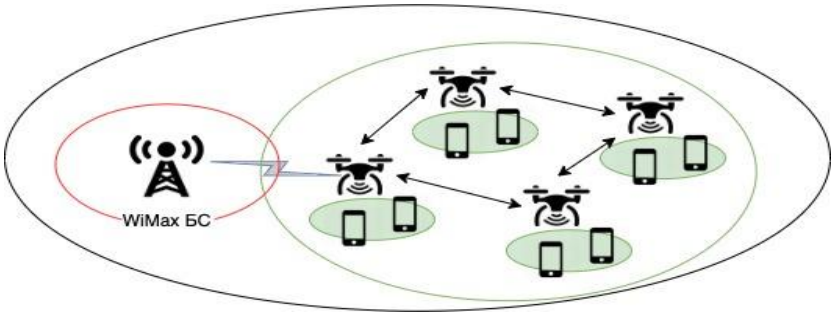



*Рисунок 3- Схема организации WiMax сети с помощью БПЛА*

При организации Wi-Fi сети с помощью БПЛА, БПЛА подключаются к сети Wi-Fi [6]. Так как зона покрытия сети Wi-Fi небольшая (до 100м), а устройство малых размеров, Wi-Fi приемо-передающее устройство может перемещаться по земле вместе с БПЛА, обеспечивая непрерывную связь (См. Рис. 4).

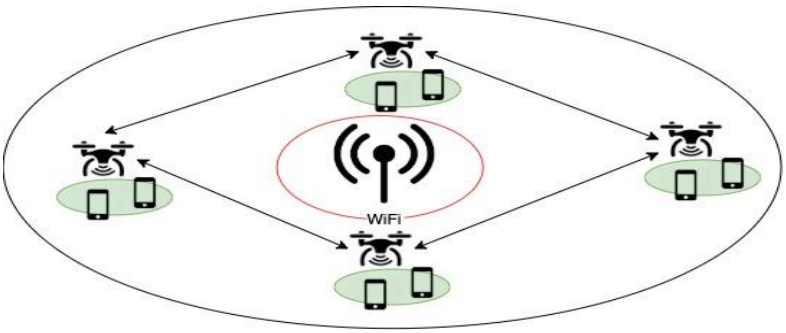



*Рисунок 4 - Схема организации Wi-Fi сети с помощью БПЛА*

При организации спутниковой сети БПЛА подключаются к существующим спутниковым сетям связи. Используются средние БПЛА 8000м, которые летают гораздо выше средних БПЛА 3000м, и покрывают связью большие зоны, ниже будут приведены советующие расчеты покрытия связью (См. Рис. 5). Такая схема рассчитана на обеспечение связью удаленных территорий, где подключение к БС невозможно или труднодоступно.

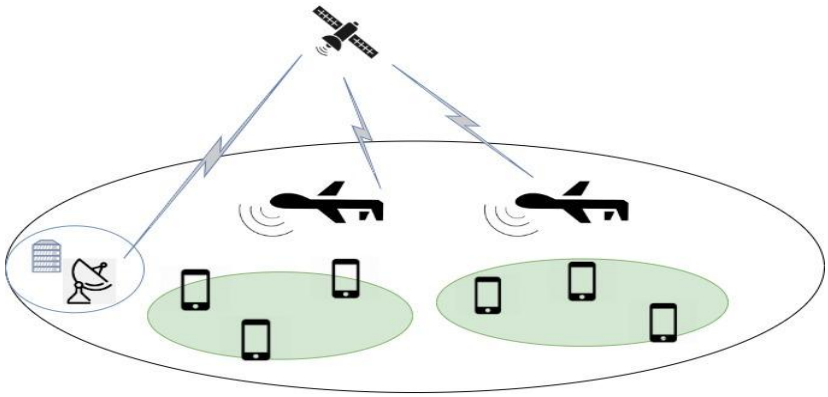


*Рисунок 5 - Схема организации спутниковой сети с помощью БПЛА*

У каждой схемы организации связи есть свои достоинства и недостатки, которые приведены в таблице 1.

Таблица 1 - Характеристики беспроводных технологий

| Схема организации связи | Достоинства | Недостатки |
|---|---|---|
| Сети сотовой связи | Легкость построения сети, низкие задержки, высокая скорость передачи данных. | Возможные помехи от поврежденных БС сотовой сети. |
| Сети WiMax связи | Скорость передачи данных, зона покрытия, возможность построения сети с помощью мобильной WiMax. | Необходимость использования соответствующих приемо-передающих устройств. |
| Сети Wi-Fi связи | Легкость построения сети, стабильность. | Малая зона покрытия. |
| Сети спутниковой связи | Дальность действия связи, зона покрытия сети. | Задержки, сложность построения сети. |

На основе применения формулы работы [10] были построены графики зависимости параметров сетей с БПЛА. Используя формулы $d_{LOS} = \sqrt{2 \cdot R_3 \cdot h_1 + h_1^2} + \sqrt{2 \cdot R_3 \cdot h_2 + h_2^2}$ и $N_{БПЛА} = S_{БС}/S_{БПЛА}$ , где

$h_1$ – высота подъема наземной антенны, м;

$h_2$ – высота подъема бортовой антенны, м;

$d_{LOS}$ – предельная дальность прямой видимости, км;

$S_{БС}$ – площадь покрытия БС, $м^2$;

$S_{БПЛА}$ – площадь покрытия БПЛА, $м^2$;

$N_{БПЛА}$ – количество БПЛА;

$R_3$ – радиус Земли (R3 = 6400 км для высоких радиочастот), были рассчитаны максимальная дальность связи с помощью БПЛА и возможное количество БПЛА, покрывающих площадь района ЧС. Таким образом, были построены график зависимости площади покрытия связи от высоты полета БПЛА и график зависимости количества БПЛА от высоты полета БПЛА.

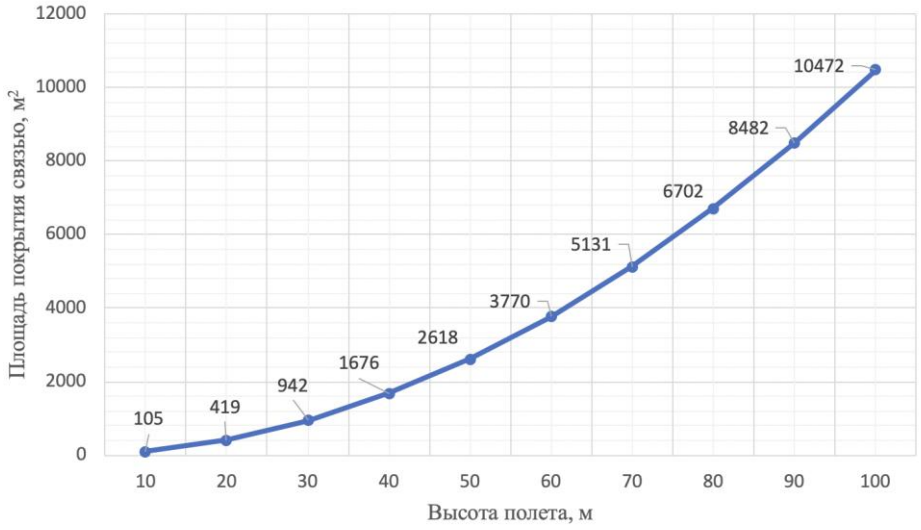



*Рисунок 6 - График зависимости высоты полета БПЛА и зоны покрытия связью*

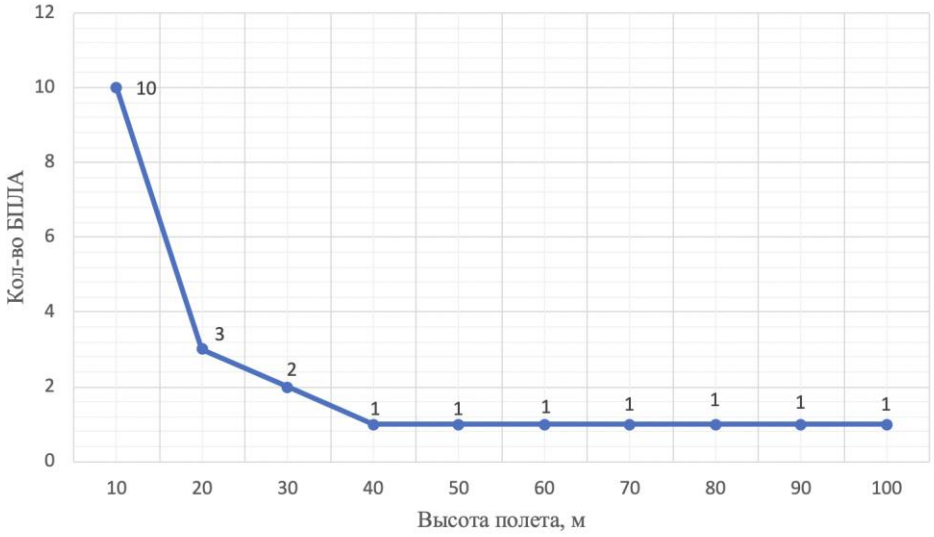



*Рисунок 7 - График зависимости количества от высоты полета БПЛА при площади покрытия связью 1000 м2*

**Заключение**

Проведенное исследование подтверждает целесообразность применения БПЛА в сетях связи при ЧС. Анализ структурных схем организации беспроводных сетей с использованием БПЛА показал возможность построения разных вариантов схем, обеспечивающих связью зону бедствия. Из детальной характеристики беспроводных технологий можно понять, в каких случаях применять определенный вид связи. Также были проведены расчеты и построены графики зависимостей параметров сетей с БПЛА, которые будут полезными при исследовании определенного района или местности и позволят сделать анализ экономической эффективности применения БПЛА в ЧС.

СПИСОК ЛИТЕРАТУРЫ

1. Zeng Y., Zhang R., & Lim T. J. (2016). Wireless communications with unmanned aerial vehicles: Opportunities and challenges. (Vol. 54). IEEE Communications Magazine. (pp. 36-42).
2. Y. Zeng Y., Zhang R., & Lim T. J. (2016). Throughput maximization for UAV-enabled mobile relaying systems. (Vol. 64). IEEE Transactions on Communications. (pp. 4983-4996).
3. Wang H., Wang J., Ding G., Chen J., Li Y., & Han Z. (2018). Spectrum Sharing Planning for Full-Duplex UAV Relaying Systems With Underlaid D2D Communications. (Vol. 36). IEEE Journal on Selected Areas in Communications. (p. 36).
4. Rahman M. A. (2014). Enabling drone communications with WiMAX Technology. DOI: 10.1109/IISA.2014.6878796. (pp. 323-328).
5. Paul N. (2021). IEEE 802 LAN/MAN Standards Committee, LMSC, LAN/MAN Standard Committeee (Project 802). [Electronic resource] URL: http://www.ieee802.org/ (accessed:20.04.2021).
6. Banerji S., & Chowdhury R. S. (2013). Wi-Fi & WiMAX: A comparative study. [Electronic resource] URL: https://arxiv.org/abs/1302.2247v11 (accessed:20.04.2021).
7. Думин Д. И., Динь Ч. З., Фам В. Д., Киричек Р. В. Применение установленных на БПЛА систем обнаружения GSM-устройств для поиска пострадавших в результате ЧС // Информационные технологии и телекоммуникации. 2018. Том 6. № 2. С. 62–69.
8. Шевцов В.А., Бородин В.В., Крылов М.А. Построение совмещённой сети сотовой связи и самоорганизующейся сети с динамической структурой // Труды МАИ. 2016. № 85. С. 4–6.

9. Ананьев А.В., Филатов С.В. Обоснование нового способа совместного применения авиации и беспилотных летательных аппаратов малой дальности в операциях // Военная мысль. 2018. № 6. С. 20–66.
10. Sathya N. V. (2013). Propagation channel model between unmanned aerial vehicles for emergency communications. [Electronic resource] URL: http://urn.fi/URN:NBN:fi:aalto-201303201831 (accessed:20.04.2021).

REFERENCES

1. Zeng Y., Zhang R., & Lim T. J. (2016). Wireless communications with unmanned aerial vehicles: Opportunities and challenges. (Vol. 54). IEEE Communications Magazine. (pp. 36-42).
2. Y. Zeng Y., Zhang R., & Lim T. J. (2016). Throughput maximization for UAV-enabled mobile relaying systems. (Vol. 64). IEEE Transactions on Communications. (pp. 4983-4996).
3. Wang H., Wang J., Ding G., Chen J., Li Y., & Han Z. (2018). Spectrum Sharing Planning for Full-Duplex UAV Relaying Systems With Underlaid D2D Communications. (Vol. 36). IEEE Journal on Selected Areas in Communications. (p. 36).
4. Rahman M. A. (2014). Enabling drone communications with WiMAX Technology. DOI: 10.1109/IISA.2014.6878796. (pp. 323-328).
5. Paul N. (2021). IEEE 802 LAN/MAN Standards Committee, LMSC, LAN/MAN Standard Committeee (Project 802). [Electronic resource] URL: http://www.ieee802.org/ (accessed:20.04.2021).
6. Banerji S., & Chowdhury R. S. (2013). Wi-Fi & WiMAX: A comparative study. [Electronic resource] URL: https://arxiv.org/abs/1302.2247v11 (accessed:20.04.2021).
7. Dumin D. I., Din' CH. Z., Fam V. D., Kirichek R. V. Primenenie ustanovlennyh na BPLA sistem obnaruzheniya GSM-ustrojstv dlya poiska postradavshih v rezul'tate CHS // Informacionnye tekhnologii i telekommunikacii. 2018. Tom 6. № 2. S. 62–69.
8. SHevcov V.A., Borodin V.V., Krylov M.A. Postroenie sovmeshchyonnoj seti sotovoj svyazi i samoorganizuyushchejsya seti s dinamicheskoj strukturoj // Trudy MAI. 2016. № 85. S. 4–6.
9. Anan'ev A.V., Filatov S.V. Obosnovanie novogo sposoba sovmestnogo primeneniya aviacii i bespilotnyh letatel'nyh apparatov maloj dal'nosti v operaciyah // Voennaya mysl'. 2018. № 6. S. 20–66.
10. Sathya N. V. (2013). Propagation channel model between unmanned aerial vehicles for emergency communications. [Electronic resource] URL: http://urn.fi/URN:NBN:fi:aalto-201303201831 (accessed:20.04.2021).

**Турбекова К.Ж.**

**Төтенше жағдайлар кезінде байланыс желілерінде ПҰА-ның қолданылуын талдау**

**Аңдатпа.** Бұл мақалада төтенше жағдайларға арналған пилотсыз ұшу аппараттарына негізделген апаттық желілердің құрылымдары талқыланған. Сымсыз технологияларға шолу жасалынған және ұшу аппараттарының көмегімен апат аймақтарында байланысты қамтамасыз ету үшін сол технологияларға негізделген желілерді ұйымдастыру схемалары ұсынылған. Сымсыз технологиялар сипаттамаларының кестесі құрылды және олардың көмегімен тәуелділік графиктері салынып, маңызды мәндеріне есептеулер жүргізілді.

**Түйінді сөздер:** пилотсыз ұшу аппараттары (ПҰА), базалық станция (БС), апаттық жағдай (АЖ), сымсыз байланыс, сымсыз технологиялар.

**Turbekova K.Zh.**

**Analysis of the use of UAVs in emergency communication networks**

**Abstract.** This article focuses on the structures of emergency networks based on UAVs for emergency situations. It presents a review of wireless technologies and schemes for the organization of networks based on such technologies to ensure communication in disaster areas using UAVs, a

table of characteristics of wireless technologies, and calculations of the major values, with which dependency charts are built.

**Keywords:** unmanned aerial vehicle (UAV), base station (BS), emergency situation (ES), wireless communication, wireless technologies.

**Автор туралы мәлімет:**

**Турбекова Карина Жанарбекқызы**, «Математикалық және компьютерлік модельдеу» кафедрасының 2-ші курс магистранты, Халықаралық ақпараттық технологиялар университеті.

**Сведение об авторе:**

**Турбекова Карина Жанарбекқызы**, магистрант 2 курса кафедры «Радиотехника, электроника и телекоммуникации», Международный университет информационных технологий.

**About the author:**

**Karina Zh. Turbekova,** master student, Department of Radio Engineering, Electronics and Telecommunication, International Information Technology University.

# ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 530.1, 681.3.06

**Azanov N.P.[1], Khabirov R.R.[2], Amirov U.E.[3]**

[1,2,3] Al-Farabi Kazakh National University, Almaty, Kazakhstan

## COMPETITIVE INTELLIGENCE AND DECISION-MAKING ALGORITHM USING MACHINE LEARNING FOR INDUSTRIAL SECURITY

**Abstract.** The purpose of this scientific article is to show what competitor data analytics can do with machine learning and neural networks. In this study, we analyzed data on potential partners of the Department of Defense Office of Hearings and Appeals (DOHA) of the USA and obtained a trained algorithm that can help in making decisions based on keywords, which can minimize reputational risks. The published dataset of the Department of Defense Office of Hearings and Appeals (DOHA) of the USA was selected for analysis of the initial data, which displayed the results of the screening of potential partners along with a text justification. This is the reason why we used Recurrent Neural Network (RNN) instead of Convolutional Neural Network (CNN). Neural networks are a very important part of machine learning. As a result, we have developed a trained machine learning model for recommending the best partners, that is, more proven partners, both professional and reputable. In addition, the developed machine learning model does not allow working with an organization of bad partners who could act in bad faith and carry reputational risks.

**Keywords**: Competitive Intelligence (CI), Data Analysis, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Machine Learning (ML).

### Introduction

If we talk about competitive intelligence (CI), we can easily find that CI is a systematic collection and analysis of information from multiple sources, as well as a coordinated CI program. This is the act of identifying, collecting, analyzing, and disseminating information about products, customers, competitors, and any aspects of the environment necessary to support managers in making strategic organizational decisions. CI means understanding and learning about what is happening in the world outside of business to improve your competitiveness, i.e. learning as much as possible about your external environment, including the industry as a whole and the relevant competitors.

Key points:

Competitive intelligence is a legal business practice, as opposed to industrial espionage, which is illegal.

The main focus is on the external business environment.

This is the process of collecting information, converting it into intelligence, and then using it in decision-making. Some CI professionals wrongly emphasize that if the information collected is not usable or actionable, it is not intelligence.

Another definition of CI sees it as an organizational function responsible for early identification of risks and opportunities in the market before they become apparent ("early signal analysis"). The term CI is often seen as synonymous with competitor analysis, but competitive intelligence is more than competitor analysis; it encompasses the entire environment and stakeholders.

### Dataset & Preparation

Finding the right dataset is not always easy. As mentioned earlier, there are more datasets in today's world that one person can handle. Nevertheless, finding the right one that fits your idea can

be challenging. We also ran into some problems while searching for data. The main issue was that some of the datasets hadn't had the features we were looking for. Suffice it to say that such a dataset could not exist without us tweaking it a little bit. And that is exactly what we did. We found a very interesting dataset about Industrial Security Clearances on Kaggle and quickly began the data cleaning process. This data contains dates, case numbers, decisions, and decision summaries for over 20,000 cases submitted for review between late 1996 and early 2016. Industry contractors that work for or with the United States Department of Defense and come into contact with secret or privileged information must submit documents for a background check by the government as a part of their contractual obligations. Any employee who fails to get the necessary clearance will not be allowed to work.

Employees may however appeal their decision; in this case the decision will be reviewed and finalized (or reversed) by the Department of Defense Office of Hearings and Appeals (DOHA). This dataset contains summaries of the deliberations and results of such hearings and provides a window into getting security clearance to work as a defense contractor in the United States.

In this research, we looked for answers to the following questions:

- What percentage of appeals were overturned or upheld?
- What percentage of appeals resulted in favorable decisions?
- What were the reasons for decisions to be overturned?
- Can a model be built to predict decisions based on the decision text i.e., an auto-classifier? [1].

We decided to analyze the criteria for selecting suppliers for defense orders for the US Department of Defense, as well as the possibility of appealing against negative results. After our analysis, conducted with the help of machine learning, we determined by what criteria you can get a favorable outcome and become one of the suppliers of defense orders.

The data consists of 5 columns: Date - date of the appeal hearing; Decision Digest - a free text field describing the keywords of the appeal; Decision Identifier - the type of leadership (e.g. financial, foreign influence, security, etc.) to which the appeal relates; Favorable Decision - a binary variable indicating the outcome of the appeal, i.e. favorable, which is rendered in favor of the party, or unfavorable, which leads to the rejection of the decision on security; Clearance Upheld - a binary variable indicating whether the decision made before the appeal was upheld by the Court of Appeal or rejected. "Yes" indicates that the judge's decision was upheld, "No" indicates that the judge's decision was overturned.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ***N*** | ***casenum*** | ***date*** | ***digest*** | ***keywords*** | ***Favorable decision*** | ***Decision upheld*** |
| 2 | 0 | 15-08250.a1 | 07/28/2017 | The Judge's adverse findin | Guideline C; Guideline B | No | Yes |
| 3 | 1 | 15-03801.a1 | 07/28/2017 | Applicant argues that the J | Guideline B | No | Yes |
| 4 | 2 | 15-07971.a1 | 07/26/2017 | Applicant's appeal brief co | Guideline F | No | Yes |
| 5 | 3 | 15-03098.a1 | 07/26/2017 | The Board cannot consider | Guideline F | No | Yes |
| 6 | 4 | 14-04693.a1 | 07/26/2017 | Applicant contends that his | Guideline F; Guideline E | No | Yes |
| 7 | 5 | 16-00844.a1 | 07/25/2017 | Applicant claims the two u | Guideline F | No | Yes |
| 8 | 6 | 15-07009.a1 | 07/25/2017 | Applicant was charged with | Guideline E | No | No |
| 9 | 7 | 15-03162.a1 | 07/25/2017 | The Judge stated that any | Guideline K; Guideline E | No | Yes |
| 10 | 8 | 16-00757.a1 | 07/24/2017 | Despite Applicant's argume | Guideline F | No | Yes |

*Figure 1 – Dataset file preview details*

Figure 1 shows the file consisting of decision text from Security Appeal hearings at the US Department of Defense. The decisions are classified by:

- Keyword – the security guideline under which clearance is sought such as Financial, Foreign influence etc.
- Favorable decision - whether the Appeals decision was favorable (security clearance granted) or unfavorable (security clearance rejected).
- Decision upheld - whether the original decision was upheld or not [1].

**Analysis (% of Yes or No)**

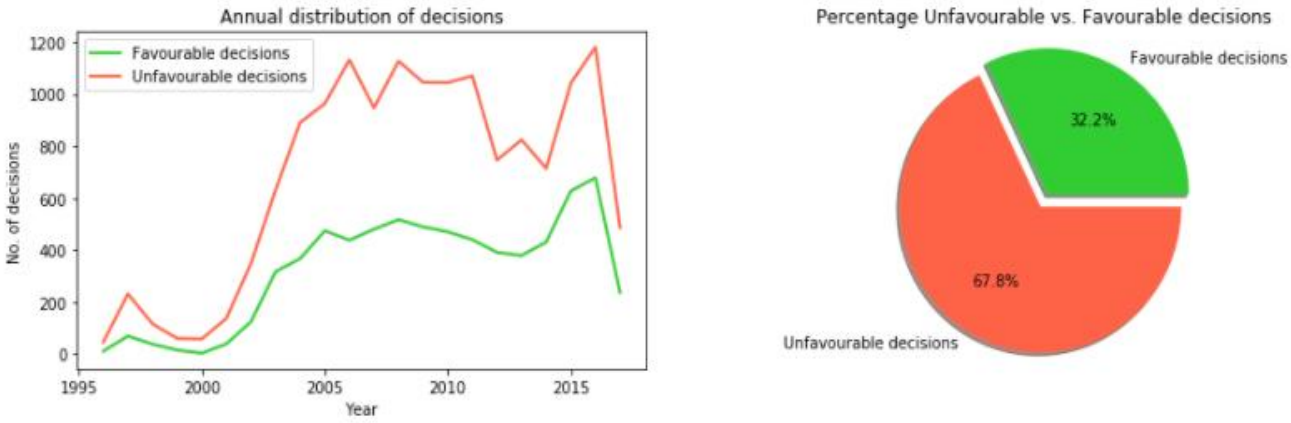



*Figure 2 - Dataset analysis*

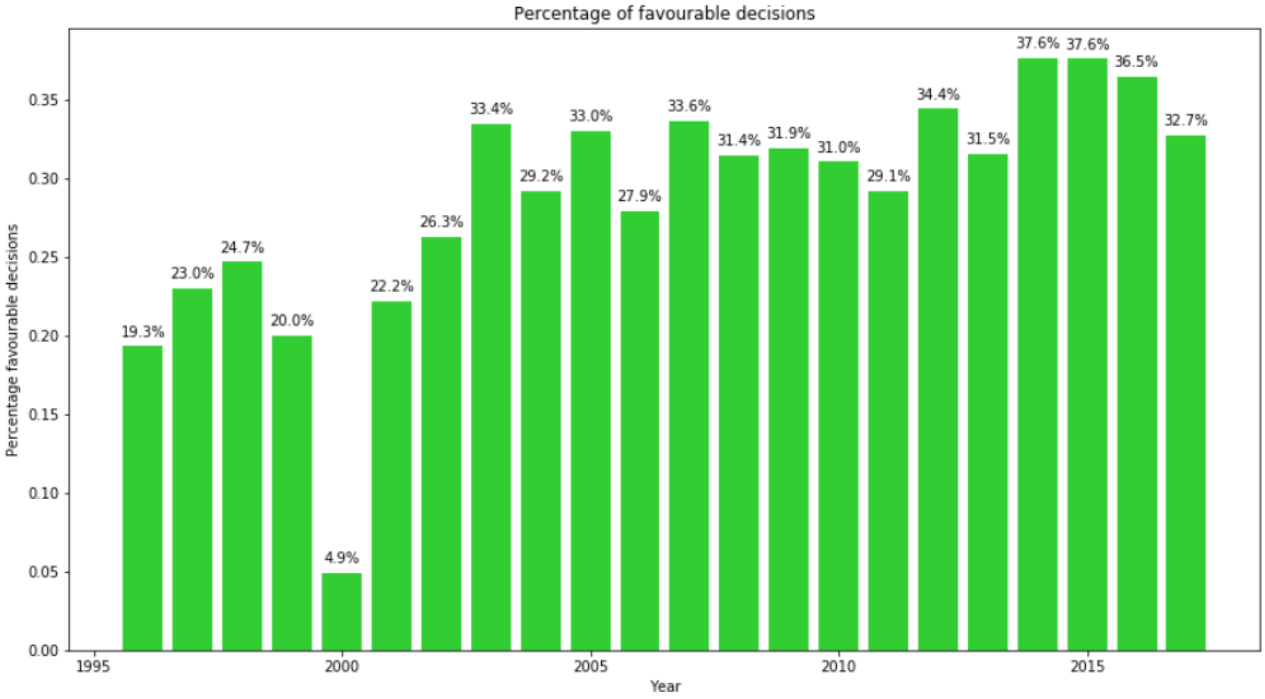



*Figure 3 - Percentage of favourable decisions*

It's evident that the majority of the cases heard by the Appeals court result in unfavorable decisions i.e. the security clearance is rejected. Almost 68% of the cases heard are rejected, which would reinforce the notion that obtaining a security clearance is a difficult task! Moreover, from the first chart, it appears that the period from 2003 onwards has seen a marked increase in the number of appeals that the court is receiving. We found, that with the increased volume of cases, the rate of rejection has increased, spiking to 37.6% in 2014 and 2015.

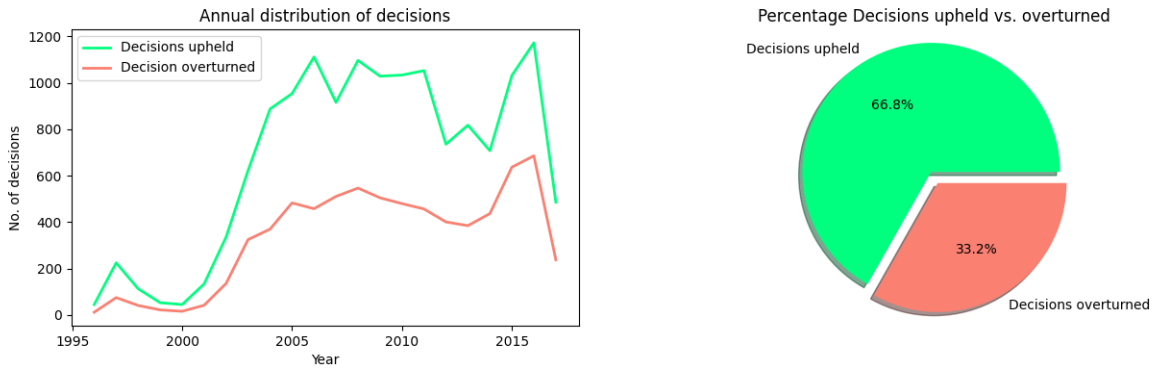



*Figure 4 - Statistics about data [1]*

Considering the increased volume and rate of rejection, it might be indicative of a more lenient process for appeals, where the court has been receiving more and more cases without merit. The subsequent set of charts looks at exploring the upheld decisions - visualization of the number upheld and overturned decisions by year, the proportion of upheld and overturned decisions, and the percentage of upheld decisions by year. Almost 1 out of 3 cases heard by the Appeals court is upheld. From the first glance at the numbers, it seems like there might be a relation between the types of cases that are upheld and the decision itself (favorable or unfavorable). Decisions overturned spiked in 2014 and 2015 at 38%, also the years in which the Appeals court received its highest volume of cases.

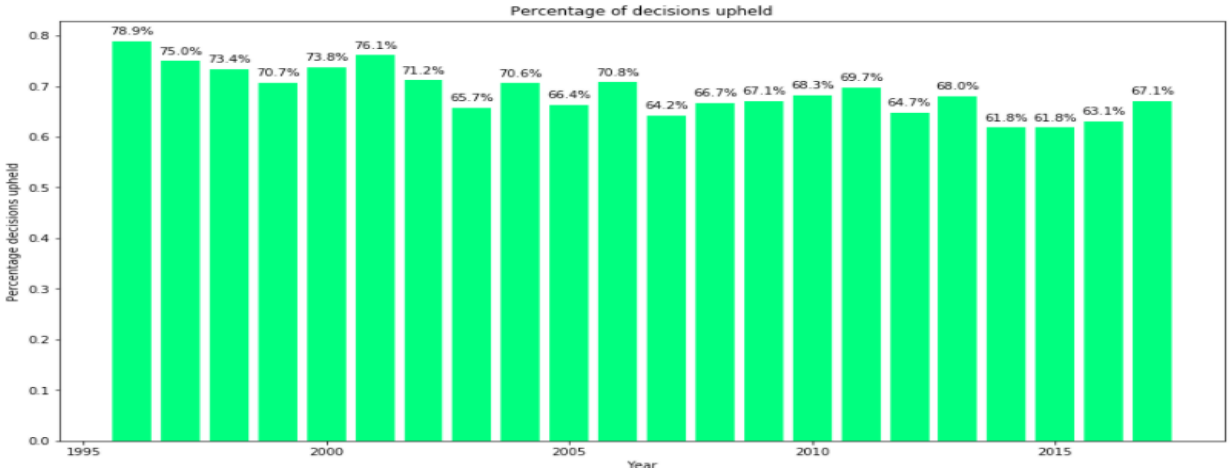



*Figure 5 - Percentage of decisions upheld*

Decisions overturned spiked in 2014 and 2015 at 38%, also the years in which the Appeals court received its highest volume of cases. It might be worth cross-tabulating overturned decisions with the outcome of the decision (favorable or unfavorable).

**ML algorithms & RNN**

In this part of the paper, we will discuss the data manipulation process which is the first step after planning that a Data Scientist undertakes. So, we followed good practice and prepared our datasets. Some of the things done include deleting the unwanted features. Some of the features were tedious and removing such redundant data gave us a better view of the important features that we were to use for our algorithm. Then dropping the entries that had missing values or fixing them where it could be done. Unfortunately, some of the entries had a lot of missing values and there was no way to predict them without losing the main goal of our project. We say this because there are a

lot of algorithms out there for fixing missing values and finding patterns to predict what those values could be, but for the purpose of simplicity, we just dropped those entries and moved on to the next step.

Neural networks are one of the most popular machine learning algorithms currently. Over time, it has been conclusively proven that neural networks are superior to other algorithms in accuracy and speed. With various options like CNNs (Convolutional Neural Networks), RNNs (Recurrent Neural Networks), AutoEncoders, Deep Learning, and so on, neural networks are slowly becoming to data scientists or machine learning practitioners what linear regression was to statisticians. Thus, it is necessary to have a fundamental understanding of what a neural network is, what it consists of, what its scope and limitations are. This article is an attempt to explain the neural network, starting with its most basic building block, the neuron, and then delving deeper into its most popular options like CNN, RNN, etc.

A neural network that has more than one hidden layer is commonly referred to as a deep neural network. Convolutional neural networks (CNN) are one of the variants of neural networks widely used in the field of computer vision [3]. The name comes from the type of hidden layers it consists of. Hidden CNN layers are usually composed of convolutional layers, merged layers, fully connected layers, and normalization layers. Here it simply means that instead of using the normal activation functions defined above, the fold and merge functions are used as the activation functions.

Recurrent neural networks, or RNNs as they are called in short, are a very important variant of neural networks widely used in natural language processing. In a conventional neural network, the input is processed through several layers, and the output is made with the assumption that two consecutive inputs are independent of each other. This assumption, however, is incorrect in a number of real-world scenarios. For example, if someone wants to predict the price of a stock at a given point in time or wants to predict the next word in a sequence, the dependency on previous observations must be considered. RNNs are called returnable because they perform the same task for each element of the sequence, with an output that depends on previous computations. Another way to think about RNNs is that they have a "memory" that collects information about what has been calculated so far [3]. In theory, RNNs can use information in arbitrarily long sequences, but in practice they are limited to looking at just a few steps. RNNs take full advantage of the context of sequential data and offer an effective and scalable model for several learning problems related to sequential data. RNNs have enabled significant progress in the fields of text processing, speech processing, and machine translation [4]. Network traffic is made up of packets, and the payloads in the packets appear in the form of data streams. When analyzing the semantics of payloads, we employ an RNN model to classify them. RNNs learn feature representations from training data by storing previous states. In the training phase, the RNN model identifies specific sequences that can distinguish normal communications from anomalous communications. RNNs can also directly process data streams; thus, RNN scans support end-to-end attack detection [4].

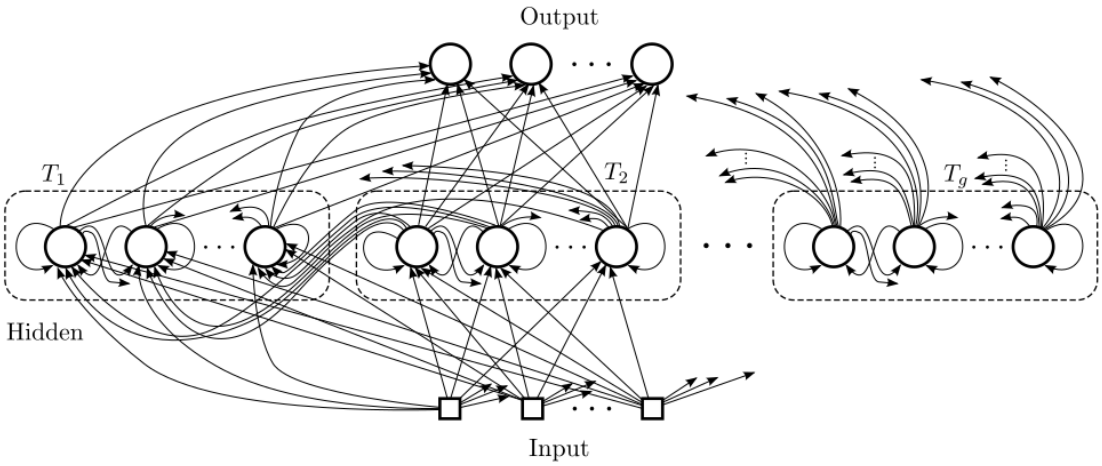



*Figure 6 - RNN with an input, output and hidden layer*

Figure 6 shows the architectural structure of the Recurrent neural network with an input layer, a couple of hidden layers and an output layer. RNNs have proven to be very successful in natural language processing, especially with their variant of LSTM, which is capable of looking back longer than RNNs.

**Implementation of RNN**

The data [1] consists of 5 columns: Date - date of the appeals hearing and decision; Digest - free text field describing the appeals decision; Keywords - identifies the type of guideline (such as Financial, Foreign Influence, Security, etc.) that the appeal relates to; Favorable Decision – a binary variable indicating the outcome of appeal i.e. favorable, which is ruled in favor of the party or unfavorable, which results in a rejection of security clearance; Decision Upheld – a binary variable indicating whether the decision made prior to appeal was upheld by the Appeals court or rejected, “Yes” indicates that the judge's decision was upheld, "No" indicates that the judge's decision was overturned; Exploring favorable decisions - visualization of the number of favorable and unfavorable decisions by year, the proportion of favorable and unfavorable decisions, and the percentage of favorable decisions by year.

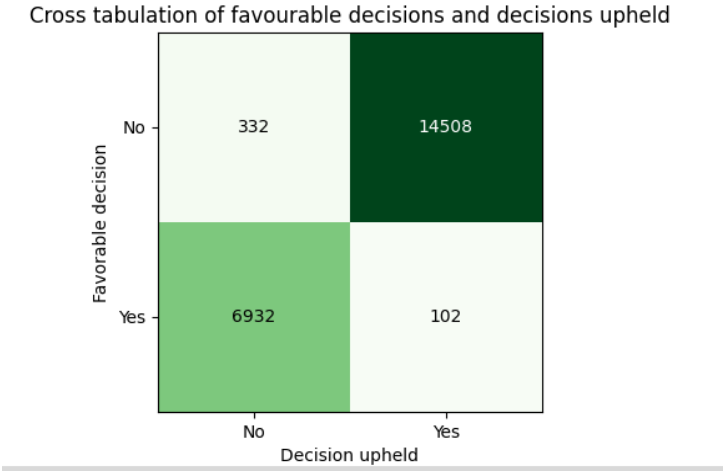



*Figure 7 - Statistics about data [1]*

Figure 7 shows cross-tabulation which confirms that out of 7,264 (6,932 + 332) decisions overturned, a majority (95% - 6,932 out of 7,264) were initially rejected applications, which were then appealed and overturned in favor of the party. Surprisingly, 5% of the appeals that were

initially ruled upon positively were later overturned in the Appeals process and security clearance was rejected. Similarly, out of 14,610 (14,508 + 102) cases that were upheld, 99% were unfavorable (security clearances were rejected). Out of 21,440 decisions that were initially ruled upon unfavorably, 32.3% of such cases (roughly 1 out of 3) were overturned and resulted in a security clearance being granted. Interestingly, an initially favorable decision that goes to the Appeals court is overturned in 76.5% of the cases.

```
48  EMBEDDING_DIM = 50
49  model = Sequential()
50  model.add(Embedding(vocab_size, EMBEDDING_DIM, input_length=max_len))
51  model.add(Bidirectional(GRU(units=32, dropout=0.15, recurrent_dropout=0.15)))
52  model.add(Dense(2, activation='sigmoid'))
53  model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
54  history = model.fit(X_train_pad, Y_train, batch_size=128, epochs=10, validation_data=(X_test_pad, Y_test), verbose=2)
```

*Figure 8 - Implementation of Bi-directional RNN with embeddings being trained with Keras*

Figure 8 presents a bi-directional RNN that can predict the outcome of whether or not the decision was upheld based on the text of the decision. With a forward RNN model, the accuracy achieved was only in the range of 60% - 70%. Bi-directional RNNs seem apt for this application since the decision text involves several double negatives and indirect references. In our implementation, we have chosen to use “sigmoid” as an activation function, because the final result must be between 0 and 1. In addition, we used “Adam” as an optimizer. Adam is an adaptive torque estimation, another optimization algorithm. It combines both the idea of motion accumulation and the idea of weaker renewal of weights for typical signs. So, at the end of calculations, the result will be demonstrated in accuracy as a metric.

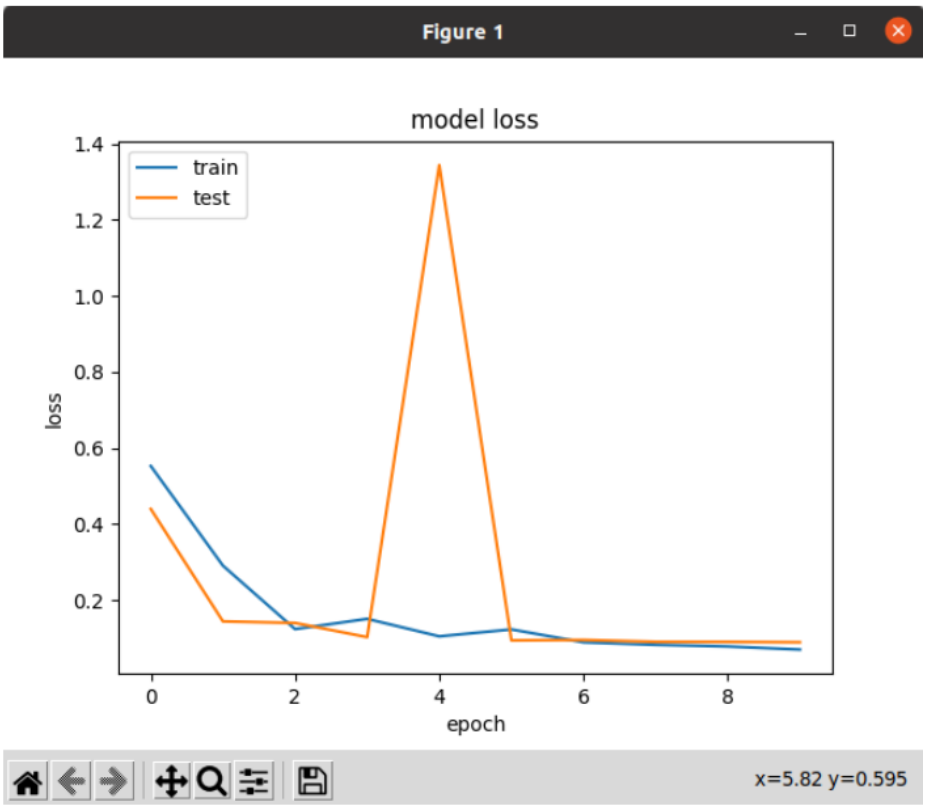



*Figure 9 - Graph of result with a comparison train and test datasets by epochs*

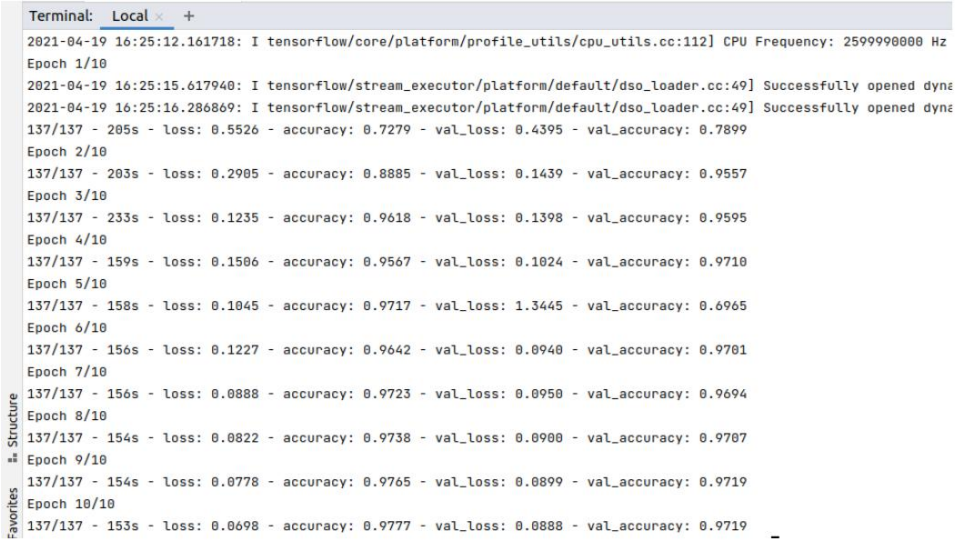



*Figure 10 - Results of each epoch in console*

Figure 10 shows the result of RRN for 10 epochs with mean **0.8129** accuracy, which is really not a bad result. It can be upgraded and finetuned in the near future by better data preparation with more efficient text analysis algorithms.

**Conclusion**

As a result, we analyzed data on the refusal or acceptance of commercial partners for cooperation with the Department of Defense Office of Hearings and Appeals (DOHA) of the USA. Classifications of the Decision Outcome (Favorable/Unfavorable decision) and Appeal Outcome (Upheld/Overturned decision) dataset have helped us understand the logical result of the decisions. In this regard, the main data is in the text form, so it was decided to use RNN since Recurrent Neural Networks are better suited for working with text than CNN, that is Convolutional Neural Networks [4].

Finally, we have implemented the RNN ML model with 0.8129 accuracy using the Python Machine Learning tools: numpy, pandas, sklearn, tensorflow and keras. The Recurrent Neural Network Machine Learning Model with high accuracy confirmed the correctness of our research. This RNN ML model can predict answers with not bad accuracy to the question: “Confirm or not confirm commercial partners to work with?” by keywords, digest, and text analysis. We can save this learned RNN ML model to reuse it in next projects or companies. We can apply our solution with this RNN ML model, already trained on this data, to aid decision-making in other ministries, enterprises, companies, etc. Based on the analyzed data, we propose a trained model for weeding out unfavorable partners to Kazakhstan organizations or government agencies.

Since CI is a process associated with collecting information, transforming it into intelligence data, in our case, using our enterprise RNN ML trained model helps commercial and government organizations in decision-making, to defend themselves from unscrupulous partners, espionage, information data leakage, and minimize the reputation risks.

REFERENCES

1. Industrial Security Clearances Classified [Electronic resource] URL: www.kaggle.com/vivek2412/industrial-security-clearances-classified (accessed:15.04.2021)
2. Yang, Jean & Hance, Travis & Austin, Thomas & Solar-Lezama, Armando & Flanagan, Cormac & Chong, Stephen. (2015). End-To-End Policy-Agnostic Security for Database-Backed Applications. [Electronic resource] URL: https://www.ftc.gov/system/files/documents/public_comments/2015/10/00039-97834.pdf

3. Arneric J, Poklepovic T, Aljinovic Z (2014) GARCH based artificial neural networks in forecasting conditional variance of stock returns. Croat Oper Res Rev 5(2):329–343
4. Hongyu Liu, Bo Lang, Ming Liu, Hanbing Yan (2019). CNN and RNN based payload classification methods for attack detection. Knowledge-Based Systems. Volume 163, 1 January 2019, Pages 332-341
5. Junxiang Fan, Qi Li, Junxiong Hou, Xiao Feng, Hamed Karimian, Shaofu Lin (2017). A Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN. [Electronic resource] URL: https://doi.org/10.5194/isprs-annals-IV-4-W2-15-2017
6. Jan Koutn´ık, Klaus Greff, Faustino Gomez, Jurgen Schmidhuber. A Clockwork RNN. IDSIA, USI&SUPSI, Manno-Lugano, CH-6928, Switzerland. Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1863-1871, 2014.
7. Susheel S., Narendra Kumar N.V., Shyamasundar R.K. (2015) Enforcing Secure Data Sharing in Web Application Development Frameworks Like Django Through Information Flow Control. In: Jajoda S., Mazumdar C. (eds) Information Systems Security. ICISS 2015. Lecture Notes in Computer Science, vol 9478. Springer, Cham. [Electronic resource] URL: https://doi.org/10.1007/978-3-319-26961-0_34.

**Азанов Н.П. Хабиров Р.Р., Әміров У.Е.**

**Конкурентная разведка и принятие решений с помощью машинного обучения для обеспечения промышленной безопасности**

**Аннотация.** Цель этой научной статьи показать, на что способна конкурентная разведка и анализ данных с помощью машинного обучения и нейронных сетей. В данном исследовании мы проанализировали данные о потенциальных партнерах Управления слушаний и апелляций Министерства обороны США (ДОХА) и получили обученный алгоритм, который может помочь в принятии решений на основе ключевых слов и который позволяет минимизировать репутационные риски. В качестве анализа исходных данных был выбран опубликованный набор данных Управления слушаний и апелляций Министерства обороны США (ДОХА), в котором наряду с текстовым обоснованием были отображены результаты скрининга потенциальных партнеров. Именно по этой причине мы использовали Рекуррентную нейронную сеть (RNN) вместо Сверточной нейронной сети (CNN). Нейронные сети - очень важная часть машинного обучения. В результате мы разработали обученную модель машинного обучения для рекомендации лучших партнеров, то есть более проверенных партнеров, как профессиональных, так и авторитетных. Кроме того, разработанная модель машинного обучения не позволяет работать организациям с неблагоприятными партнерами, которые могут действовать недобросовестно и нести репутационные риски.

**Ключевые слова**: Конкурентная разведка, Анализ данных, Рекуррентная нейронная сеть (RNN), Сверточная нейронная сеть (CNN), Машинное обучение.

**Азанов Н.П. Хабиров Р.Р., Әміров У.Е.**

**Өнеркәсіптік қауіпсіздікті қамтамасыз ету үшін машиналық оқытуды қолдана отырып, бәсекеге қабілеттілікті барлау және шешім қабылдау**

**Аңдатпа**. Бұл ғылыми мақаланың мақсаты машиналық оқыту мен нейрондық желілерді қолдана отырып, бәсекеге қабілетті барлау мен деректерді талдауға қабілетті екенін көрсету. Бұл зерттеуде біз АҚШ Қорғаныс министрлігінің (ДОХА) тыңдаулар мен апелляциялар бөлімінің әлеуетті серіктестері туралы деректерді талдадық және кілт сөздерге негізделген шешім қабылдауға көмектесетін және беделді тәуекелдерді азайтуға мүмкіндік беретін оқытылған алгоритм алдық. Бастапқы деректерді талдау ретінде АҚШ Қорғаныс министрлігі (ДОХА) тыңдаулар мен апелляциялар Басқармасының жарияланған мәлімет жиынтығы таңдалды, онда мәтіндік негіздememен қатар әлеуетті серіктестердің скринингтік нәтижелері көрсетілді. Дәл осы себепті біз конвульсиялық нейрондық желінің (CNN) орнына

қайталанатын нейрондық желіні (RNN) қолдандық. Нейрондық желілер машиналық оқытудың өте маңызды бөлігі болып табылады. Нәтижесінде біз ең таңдаулы серіктестерді, яғни кәсіби және беделді серіктестерді ұсыну үшін машиналық оқытудың оқытылған моделін жасадық. Сонымен қатар машиналық оқытудың дамыған моделі жосықсыз әрекет ете алатын және беделді қауіп-қатерге душар болатын қолайсыз серіктестермен ұйымдарға жұмыс істеуге мүмкіндік бермейді.

**Түйінді сөздер**: Бәсекеге қабілетті барлау, деректерді талдау, қайталанатын нейрондық желі (RNN), конвульсиялық нейрондық желі (CNN), машиналық оқыту.

**Авторлар туралы мәлімет:**

**Азанов Николай Прокопьевич**, физика-математика ғылымдарының кандидаты, әл-Фараби атындағы Қазақ Ұлттық университеті доценті.

**Хабиров Роман Рамильевич**, әл-Фараби атындағы Қазақ Ұлттық университеті магистранты.

**Әміров Уалихан Ержанулы**, әл-Фараби атындағы Қазақ Ұлттық университеті магистранты.

**Сведения об авторах:**

**Азанов Николай Прокопьевич**, канд. ф.м. наук, доцент, Казахский национальный университет им. Аль-Фараби, Алматы, Казахстан.

**Хабиров Роман Рамильевич**, магистрант, Казахский национальный университет им. Аль-Фараби.

**Әміров Уалихан Ержанулы**, магистрант, Казахский национальный университет им. Аль-Фараби.

**About the authors:**

**Nikolai P. Azanov,** Cand. Sc. **(**Ph.Math.**)**, Associate Professor, Al-Farabi Kazakh National University.

**Roman R. Khabirov,** master student, Al-Farabi Kazakh National University.

**Amirov E. Ualikhan,** master student, Al-Farabi Kazakh National University.

УДК 530.1, 681.3.06

**Janybekova S.T.[1], Tolganbayeva G.A.[2], Sarsembayev A.A.[3]**

[1,2,3] International Information Technology University, Almaty, Kazakhstan

# SPEAKER RECOGNITION USING DEEP LEARNING

**Abstract.** This paper discusses a transition from the traditional methods to novel deep learning architectures for speaker recognition. The article aims to compare the traditional statistical methods and new approaches using deep learning models. To articulate the difference in the discussed approaches it furthermore describes several recent methods of optimization and evaluation techniques. The review covers datasets used, results, contributions made toward speaker recognition, and limitations related to it.

**Keywords:** Speaker recognition, convolutional neural network, deep neural network, voice identification, recognition systems.

## Introduction

Speaker recognition is the process of voice identification among a given set of speakers from a speech signal. Speech signal conveys information about the speaker's physiological properties [2], as well as behavioral aspects like accent and involuntary transforms of acoustic parameters [1] and several widely known application domains of speaker recognition. A user authentication in the bank sector is one of these applications. In February 2016 UK high-street bank HSBC [11] and its internet-based retail bank First Direct announced that it will offer its biometric banking software to access online and phone accounts using their fingerprint or voice [4] to their 15 million customers. Another application domain of growing popularity is home assistance.

The field of speaker recognition can be divided into speaker identification and speaker verification. It may be either open-set or closed-set [5]. The aim of speaker identification is to determine whether the voice of an unknown speaker matches one of the other speakers in a dataset, where the number of speakers could be very large. Speaker verification, on the other hand, is the process of determining who is speaking from known voices in the database. If the population of recorded voices is fixed, it is an open set. In contrast, closed-set verification is a case when new recordings of people can be added without having to redesign the system [6].

There are two types of systems commonly used in speaker recognition: text-dependent and text-independent. Text-dependent systems understand the content of the speech. Usually, the content is short utterances. In text-independent systems, there is no restriction on the spoken text. Forensic speaker ID is an example of text-independent applications [7].

Later in this article, we will discuss common traditional systems of speaker recognition such as GMM-UBM, GMM-SVM, and their limitations, as well as the use of deep learning technologies that have significantly advanced speaker recognition performance.

## Feature extraction

The speech sound is a time-variant expressing various types of information, including text, speaker identities, acoustic features, emotions, etc. Speech may be observed in the time domain and frequency domain [4]. The most commonly used tool for visualizing speech is a spectrogram which describes the frequency spectra of consecutive short-term speech segments as an image. In the image, the horizontal and vertical axes represent time and frequency. The concentration of each point in the image is the magnitude of a distinct frequency and time. But for statistical modeling, spectrograms are not the best way [4]. One of the reasons is that the frequency dimension is too high. For the 1024-point fast Fourier transform, the frequency dimension is 512. Another reason is high correlation of frequency components with each other after FFT. A more compact illustration of

speech is obtained by using cepstral representation. Mel-frequency cepstral coefficients (MFCCs) are used to get features of speech. Figure 1.1 shows the technique of extracting MFCCs from a frame of the speech signal. In the figure, s(n) corresponds to a frame of speech, X(m) is the logarithm of the spectrum at frequencies determined by the *m*th filter in the filter bank, and MFCCs.

$$o_i = \sum_{m=1}^{M} \cos\left[i\left(m - \frac{1}{2}\frac{\pi}{M}\right)\right] X(m), i = 1\ldots, P \quad (1.1)$$

$$e = [e, o_1., o_p, \Delta e, \Delta o_1.., .\Delta o_p, \Delta\Delta e, \Delta\Delta o_1..\Delta\Delta o_p]^T \quad (1.2)$$

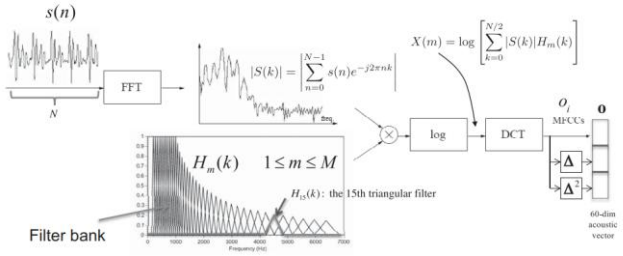



*Figure 1.1 - Procedure of extracting MFCCs from a frame of speech. Refer to Eq. 1.1 and Eq. 1.2 for $o_i$and o respectively [4]*

In Figure 1.2, the symbols $\Delta$ and $\Delta\Delta$ represent the velocity and acceleration of MFCCs, respectively. Denoted as $e$ is the log-energy, an acoustic vector corresponding to $s(n)$.

One of the widespread approaches is the use of short-term spectral features: Mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), linear predictive cepstral coefficients [1]. They are used due to their high performance and relatively low computational complexity [1].

Some recent DNN works suggest speaker recognition from raw waveforms using Convolutional Neural Networks (CNNs). Instead of using standard hand-crafted features, networks learn low-level speech representations from raw waveforms. That allows to better get significant narrow-band characteristics such as pitch and formats. ResNet-base, VGG-M based trunk CNN architectures are used for spectrogram inputs [5].

**Probabilistic Models**

One category of statistical learning methods is distinguished as latent variable models that intend to relate a set of observable variables X to a set of latent variables Z based on the number of probability distributions. $\Lambda$ parameters of the probability functions are evaluated by maximizing the likelihood function with respect to model parameters [5].

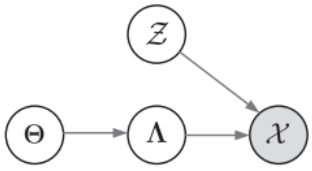



*Figure 2.1 - A hierarchical Bayesian representation for a latent variable model with parameters, latent variables $Z$, observed variables $\chi$ and hyperparameters. The shaded node denotes observations. The unshaded nodes mean latent variables [5]*

Based upon the hyperparameters , model parameters are represented by a prior density $p(\Lambda|\Theta)$ . Having the model parameters, the speech training samples $\chi$ are generated by a likelihood

function $p(\chi|\ \Lambda)$, which is marginalized over discrete latent variables by

$$p(\chi\ |\Lambda)\ =\ \Sigma_z\quad\ p\ (\chi|\Lambda) \tag{1.3}$$

through a continuous latent variable by

$$p(\chi\ |\Lambda)\ =\ \int\quad\ p\,(\chi, Z\ |\ \Lambda)\ dZ \tag{1.4}$$

The conditional likelihood with discrete latent variables is expressed by

$$p\ (Y|\ \chi,\Lambda)\ =\ \Sigma_z\quad p(\ Y,\Lambda)\ =\ \Sigma_z\quad\ p(Y|\chi,\Lambda).p(Z|\Lambda) \tag{1.5}$$

Unsupervised, supervised, and semi-supervised models can be adaptably performed and constructed by optimizing the corresponding likelihood functions in an individual or hybrid style [5].

The probabilistic methods are complex for real-world applications. There is a variety of approximate inference algorithms that can solve the optimization problem, but they are generally hard to derive. Indirect optimization over the evidence lower bound (ELBO) is implemented as an analytical solution [5]. There are different latent variable models such as Gaussian Mixture models (GMM), joint factor analysis (JFA), probabilistic linear discriminant analysis (PLDA), factor analysis (FA), a mixture of PLDA. The GMM-UBM system is a direct generative approach for speaker verification tasks. In this method, the training phase is preceded by the estimation of s speaker-independent universal background model (UBM), using a large voice data of several hours. The UBM is where

$$\lambda\ _{UBM}\ =\ \{w_i,\mu\ _i,\Sigma\ _i\}\ _{i=1}^{C} \tag{1.6}$$

C is the quantity of Gaussian components, $\omega\ _i$ is the prior of $i$-th Gaussian component, $\mu\ _i$ is mean and $\Sigma\ _i$ - covariance matrix. Each speaker is an adaptation from UBM.

i-vector system is the high dimensional GMM supervector in a total variability (TV) space. It reduces the supervector into low dimensional factors [1]. Concatenated means of GMM are presented as

$$M\ =\ m\ +\ \Phi y \tag{1.7}$$

where $\Phi$ is a low-rank factor loading matrix and $m$ - channel. $M$is a supervector of speech utterance with feature vectors.

**Deep Neural Networks**

Deep Neural Networks are trained to discriminate between speakers, map variable-length utterances to fixed-dimensional embeddings that are called x-vectors. Although most speaker recognition systems are based on i-vectors, x-vectors used like i-vectors bit built on DNN embedding architectures [6] are used in novel publications. For visual comparison there are i-vector and x-vector pipelines shown in Figure 3.1

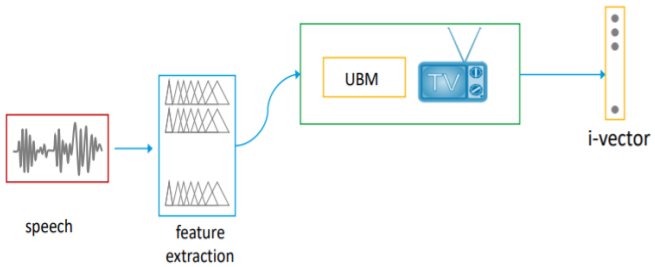



*Figure 3.1 - The i-vector pipeline [7]*

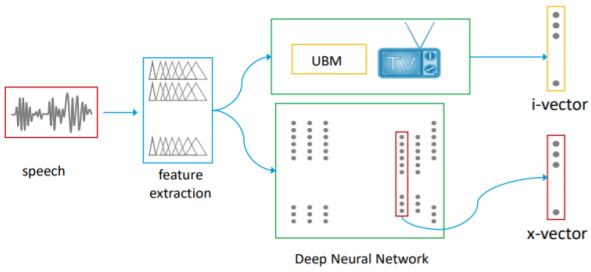



*Figure 3.2 - The i-vector and x-vector pipelines [7]*

The x-vector system is based on a framework for speaker recognition. The system is composed of a feed-forward deep neural network that maps variable-length speech segments to embeddings that are called x-vectors [8]. Those vectors are classified by trained Gaussian classifiers. The network is implemented using the nnet3 neural network library in the Kaldi Speech Recognition Toolkit [3]. The recipe is based on the SRE16 v2 recipe available in the main branch of Kaldi [8].

The *d*-vector was developed using multiple fully-connected neural network layers and *X*-vectors which are based on the Time-delayed neural networks (TDNN) that are popular in recent years [7].

**Evaluation metrics**

For closed-set speaker recognition, accuracy (recognition rate) is the usual performance measure [4].

$$Recognition\ rate = \frac{\#\ of\ correct\ recognition}{Total\ \#\ of\ trials} \tag{1.8}$$

Other popular measures include the false rejection rate (FRR), the actual decision cost function (DCF), the minimum decision cost function (min DCF), and the equal error rate (EER). Their principles are very similar.

The definition of FAR and FRR are the following:

$$False\ reject\ rate\ (FRR) = Miss\ probability = \frac{\#\ of\ true-speakers\ rejected}{Total\ \#.of\ true-speaker\ trials} \tag{1.9}$$

$$False\ acceptance\ rate\ (FAR) = \frac{\#\ of\ impostors\ accepted}{Total\ \#\ of\ impostors\ attempts} \tag{2.1}$$

Concepts of FAR, FRR, ERR are explained in Figure 4.1. They use the distributions of speaker scores and impostor scores of two speaker verification systems which are System A and System B.

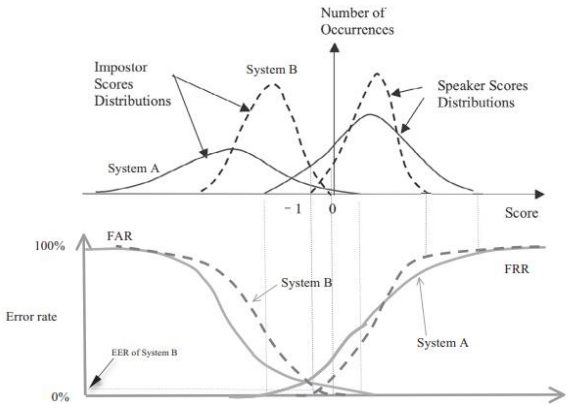



*Figure 4.1 - Distributions of true speaker scores and impostor scores of two speaker verification systems. EER, FAR, FRR of two systems [4]*

Some results of the comparison and advancements on i-vector based and x-vectors speaker recognition techniques are presented in Table 1.4 for better understanding.

The primary performance measure for the Conversational Telephone Speech (CTS) Speaker Recognition Challenge (CTS SRE) in 2019 was a detection cost described as a weighted sum of false-reject (miss) and false-accept (false alarm) error probabilities. The CTS Challenge primarily normalized the cost function for a decision threshold θ [3]

$$C_{norm}(\Theta) = P_{miss}(\Theta) + \beta \times P_{fa}(\Theta), \quad (2.2)$$

where $\beta$is

$$\beta = \frac{C_{fa}}{C_{miss}} \times \frac{1 - P_{target}}{P_{target}} \quad (2.3)$$

where $C_{min}$ - is the cost of a missed detection and $C_{fa}$ - the cost of a false alarm. $P_{target}$is the a priori probability that the test segment speaker is the specified target speaker. The primary cost metric, $C_{primary}$ for the CTS Challenge was the average of normalized costs calculated at two points along the detection error trade-off (DET) curve [3], with $C_{miss}$=$C_{fa}$= 1, $P_{target}$ = 0.01, and $P_{target}$ = 0.005.

Table 1.1 Recent results of EER i-vectors and DNN methods

| Year | Datasets | Evaluation Results | |
|---|---|---|---|
| 2018<br>Results for VoxCeleb 1 verification [9] | VoxCeleb 1 consists of over 100,000 utterances for 1,251 celebrities, extracted from videos uploaded to YouTube [9]. | GMM-UBM | 15.0 |
| | | I-vectors + PLDA | 8.8 |
| | | CNN - 1024 | 10.2 |
| | | CNN + Embedding | 7.8 |
| 2018<br>Results for verification on the original VoxCeleb2 test set [10] | Test VoxCeleb2<br>VoxCeleb2 consists of over 1 million utterances for over 6,000 celebrities. The dataset is reasonably gender-balanced, where 61% of the speakers are males | VGG-M | 5.94 |
| | | ResNet-34 | 4.83 |
| | | ResNet-50 | 3.95 |
| 2020<br>Verification performance on full utterance (NS: Normalized softmax; TAP: Temporal Average Pooling) [12] | VoxCeleb 1 | ResNet34 (TAP+NS) | 3.81 |
| | VoxCeleb 2 | ResNet34 (TAP+NS) | 2.08 |
| 2020<br>E-TDNN (Densely Connected Time Delay Neural Network for | VoxCeleb1 test set | E-TDNN | 4.65 |

| | | | |
|---|---|---|---|
| Speaker Verification) [13] | | | |
| 2020 Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs [13] | VoxCeleb1 test set | D-TDNN-SS | 1.22 |

**Conclusion**

This paper has provided a brief review of Probabilistic Models and deep learning techniques for speaker recognition. Deep learning techniques such as CNN have been the subject of much research in recent years. This research considers limitations of traditional techniques and forms a base to evaluate the performance and limitations of the current deep learning techniques. Further, it highlights some promising directions for better speaker recognition systems. It is still a non-trivial task when recognizing or verifying speakers in poor acoustic conditions. The paper also compares several model deep learning approaches for speaker recognition tasks, their evaluation metrics and datasets.

REFERENCES


1. Poddar, A., Sahidullah, M. and Saha, G., 2017. Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics, 7(2), pp.91-101.
2. Kinnunen, T. and Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), pp.12-40.
3. Sadjadi, S.O., Greenberg, C., Singer, E., Reynolds, D., Mason, L. and Hernandez-Cordero, J., 2020, May. The 2019 NIST speaker recognition evaluation CTS challenge. In Speaker Odyssey (Vol. 2020, pp. 266-272).
4. Mak, M.W. and Chien, J.T., 2020. Machine learning for speaker recognition. Cambridge University Press pp. 3-72.
5. Ravanelli, M. and Bengio, Y., 2018, December. Speaker recognition from raw waveform with sincnet. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 1021-1028). IEEE.
6. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S., 2018, April. X-vectors: Robust dnn embeddings for speaker recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5329-5333). IEEE.
7. Kelly, F., Alexander, A., Forth, O. and van der Vloed, D., From i-vectors to x-vectors–a generational change in speaker recognition illustrated on the NFI-FRIDA database, pp. 1-5.
8. Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D. and Khudanpur, S., 2018, June. Spoken language recognition using x-vectors. In Odyssey (pp. 105-111).
9. Nagrani, A., Chung, J.S. and Zisserman, A., 2017. Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612. (pp. 1-3).
10. Chung, J.S., Nagrani, A. and Zisserman, A., 2018. Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (pp. 1-5).
11. HSBC Holdings plc is a British multinational investment bank and financial services holding company, https://en.wikipedia.org/wiki/HSBC.
12. Kye, S.M., Jung, Y., Lee, H.B., Hwang, S.J. and Kim, H., 2020. Meta-learning for short utterance speaker recognition with imbalance length pairs. arXiv preprint arXiv:2004.02863 (pp. 1-5).
13. Yu, Y.Q. and Li, W.J., 2020. Densely Connected Time Delay Neural Network for Speaker Verification. Proc. Interspeech 2020, (pp.921-925)

**Джаныбекова С.Т., Толғанбаева Г.А., Сарсембаев А.А.**
**Распознавание говорящего с помощью глубокого обучения**

**Аннотация:** В этой статье обсуждается переход от традиционных методов к новым архитектурам глубокого обучения для распознавания говорящего. Он направлен на сравнение традиционных статистических методов и новых подходов с использованием моделей глубокого обучения. Также описаны новейшие методы оптимизации. Из-за разных подходов существует несколько методик оценки. В этой статье представлен обзор методов глубокого обучения и обсуждается недавняя литература, в которой эти методы используются для распознавания речи. Обзор охватывает используемые базы данных, результаты, вклад в распознавание речи и связанные с этим ограничения.

**Ключевые слова:** распознавание говорящего, сверточная нейронная сеть, глубокая нейронная сеть, идентификация по голосу, системы распознавания.

**Джаныбекова С.Т., Толғанбаева Г.А., Сарсембаев А.А.**
**Терең оқыту арқылы сөйлеушіні тану**

**Аңдатпа:** Бұл мақалада сөйлеушілерді тану үшін дәстүрлі әдістерден жаңа терең оқыту архитектурасына көшу туралы айтылады. Ол тереңдетілген оқыту модельдерін қолдана отырып дәстүрлі статистикалық әдістер мен жаңа тәсілдерді салыстыруға бағытталған. Сонымен қатар оңтайландырудың соңғы әдістері сипатталған. Сондай-ақ әртүрлі тәсілдерге байланысты бағалаудың бірнеше әдістемесі бар. Бұл мақалада терең оқыту әдістеріне шолу жасалады және сөйлеуді тану үшін осы тәсілдерді қолданатын соңғы әдебиеттер талқыланады. Шолу пайдаланылған мәлімет базасын, нәтижелерді, сөйлеуді тануға қосқан үлестерін және осыған байланысты шектеулерді қамтиды.

**Түйінді сөздер:** сөйлеушіні тану, конволюциялық жүйке жүйесі, терең жүйке жүйесі, дауысты сәйкестендіру, тану жүйелері.

**Авторлар туралы мәлімет:**

**Сарсембаев Айдос Айдарович,** PhD «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының ассистенті, Халықаралық ақпараттық технологиялар университеті.

**Джаныбекова Салтанат Талгатбековна**, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының докторанты, Халықаралық ақпараттық технологиялар университеті.

**Толғанбаева Гауһартас Алғабасқызы** «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының докторанты**,** Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Сарсембаев Айдос Айдарович,** PhD, ассистент-профессор кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**Джаныбекова Салтанат Талгатбековна,** докторант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**Толғанбаева Гауһартас Алғабасқызы,** докторант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**About the authors:**

**Aidos A. Sarsembayev**, Ph.D., Assistant-Professor, Department of Computer Engineering and Information Security, International Information Technology University.

**Saltanat T. Janybekova** doctoral student, Department of Computer Engineering and Information Security, International Information Technology University.

**Gaukhartas A. Tolganbayeva**, doctoral student, Department of Computer Engineering and Information Security, International Information Technology University.

УДК 004.352.242

**Salerova D.K. [1], Sarsembayev A.A.[2]**

[1,2] International Information Technology University, Almaty, Kazakhstan

# REVIEW OF LICENSE PLATE RECOGNITION USING OPTICAL CHARACTER RECOGNITION

**Abstract.** With the development of technologies and population increase the need in License Plate Recognition systems becomes more and more urgent. The police forces, park management, traffic control employees use LPR to improve their work. LPR system captures the image from digital camera, then makes pre-processing, character segmentation and character recognition. This paper is the survey of the OCR techniques and phases, the LPR application areas with a brief description of algorithms and methods applied by other researchers.

**Keywords:** License Plate Recognition (LPR), Optical Character Recognition (OCR), ANPR (Automatic Number Plate Recognition), OCR algorithms, CER (Character Error Rate).

## Introduction

Optical Character Recognition (OCR) is the data mining of text from image and then transforming the text for editing or searching.

License Plate Recognition (LPR) is an image processing mechanism used to recognize vehicles by their license plates. LPR was invented in 1976 by the UK Police Research Branch. The first test systems were deployed on the A1 road and in the Dartford tunnel. The first arrest on finding a stolen car was made in 1981 [1]. Creation of an effective LPR system is one of the most dynamic research spheres that has such applications as traffic control and optimization, car park management, border control and law enforcement.

The detection and extraction of data from a number plate is the most important part of the LPR system that influences its overall accuracy. The performance can be hindered by the presence of noise, rough illumination, blurring, foggy conditions and unclear light. Most LPR algorithms usually consist of four steps: image capture, number plate detection, character segmentation and recognition. Character segmentation methods depends on the input image size, various lightening conditions, license plate location. Therefore, it is irrational to exactly conclude which method really shows the highest performance.

## Literature review

The section focuses on the works previously done by several researchers. Optical character recognition is a broad topic that has its own history. In 1929, almost a century ago, Gustav Tauschek received a patent for an OCR method in Germany. The Tauschek machine was a mechanical device that used templates and a photo detector. The problem of automatic vehicle number plate recognition has been explored since the mid 90’s.

## Survey on OCR

The works of Sarika Pansare et al. [2] and Sukhpreet Singh [3] familiarize with approaches used for the design of OCR systems. Venkata Rao et al. [4] give the comparison techniques and explain the phases of OCR. Noman Islam et al. in their work [5] demonstrate the types of OCR systems: handwriting and machine printed character recognition. The table of major phases of OCR gives the description and approaches used in OCR which has a wide sphere of applications [4][5] such as banking, medicine, transportation, etc.

**Algorithms**

Most articles in optical character recognition on license plate use Convolutional Neural Network to extract features and recognize the vehicle number.

Michael Reynaldo Phangtriastu [6] describes the difference between neural network and support vector machine in OCR by comparing some feature extraction techniques and their combination.

Joseph Tarigan [7] tries to define the performance of Genetic Algorithms in improving the learning rate, the number of hidden neurons and the momentum rate on Back Propagation Neural Network.

Anuja P. Nagare [8] uses Back Propagation and Learning Vector Quantization techniques for character recognition of license plate image.

The table of Chirag Patel in [9] gives a comprehensive description of 15 works featuring OCR methods, their type and success rate. The table clearly presents the number plate detection rate and image size from several papers.

The work of Narendra Sahu et al. [10] reviews the performance examination of optical alphabet identification and also briefly explains the manuscript processing steps and their potential difficulties.

The paper of Nur- A- Alam et al. [11] shows the method for detecting and recognizing vehicle number plates in Bangladesh. The images of the vehicles are caught, and the number plate districts are recovered by the template matching method. After segmentation, CNN recover the features of each character categorizing the vehicle city, type and number.

The work of S ́ergio Montazzolli Silva [12] proposes the Automatic License Plate Recognition system and uses 5 different datasets to show the results.

Stuti Asthana et al. in their work [13] describe the back-propagation algorithm using two hidden layers that provides for high accuracy.

Humayun Karim Sulehria et al. [14] apply such methods as Back-Propagation Network, Feed-Forward Neural Network, Maximum Entropy and Hough Transform, and compare them to the proposed method.

Two tables in [15] from Divya Gilly accurately describe and compare different License Plate Recognition systems and License Plate Detection systems.

**Phases of OCR**

Optical character recognition is the complex process of actions that are usually called phases. According to Figure 1 they are:

Image acquisition - the digital image of the text is collected using a scanner or a camera, which can be subject to digitization, binarization and compression.

Preprocessing – a set of techniques such as thresholding and noise removal, used to improve the quality of the image. It is important because its defects could affect the recognition rates.

Character segmentation - breaking a segment image into component characters. There are two types of segmentation: explicit and implicit.

Feature extraction – identification of different features from characters, following segmentation. These features help to identify the character.

Character classification - the process of relating the character features to a corresponding class. The character classification techniques may be statistical, structural, syntactic.

Post processing is done to perfect the accuracy of an optical recognition system. Natural language processing, linguistic and geometric contexts are applied to fix errors.

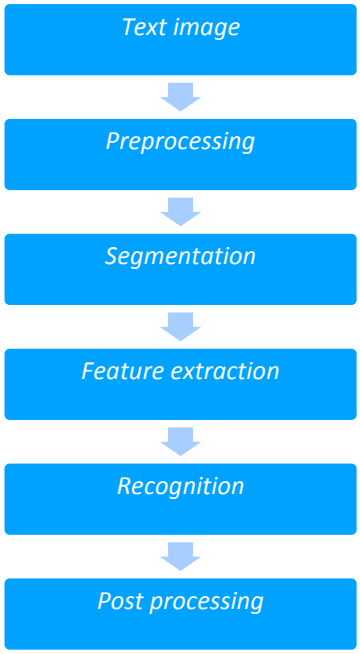



*Figure 1 - Main phases of OCR*

**OCR techniques**

Matrix matching

Matrix matching interprets a character within a pattern and compares it with a library of character templates.

Fuzzy logic

A fuzzy system receives data input from sensors and concludes findings based on it. Findings usually serve as the base for a control system.

Feature extraction

The aim of feature extraction is to select such features that increase the recognition rate. Here each character is introduced like a vector of features.

Structural analysis

Structural analysis recognizes characters by checking the horizontal, sub feature, and sub vertical forms of the image.

Neural networks

This technology simulates the neural network of the human brain. A neural network tests the pixels of the image and relates them to the existing index of character patterns.

**Application in LPR**

Automatic License Plate Recognition ensures automated access to the number plates for database systems that contain data about vehicle movements.

Parking

Automation and security include payment for parking without a ticket, car theft prevention, parking access automation, determining the car location.

Access Control

The permission of access to resources based on the users' personalities and fellowship in different groups. License plate recognition provides automation access to vehicle control with enhanced security, car park administration for logistics and events control.

Traffic Control and Optimization

License plate recognition is especially applied to detect the speed of a moving vehicle. LPR notes the average speed of cars and mitigates the hurry during traffic jams.

Border Control

Represents the country's border security from illegal cross border traffic, contraband and terrorism. It helps to decrease the level of violent crime and raise the social security.

Law Enforcement

License plate recognition contributes to quick recognition of stolen cars based on current backlist, issuance of fines for red light passage and overspeeding.

**Comparison of studied works**

Table 1 below presents metrics from other researchers' work with their progress. It represents character error rate (CER), character recognition rate (CRR) and accuracy.

Table 1 - Metrics comparison

| Works | CER / CRR | Accuracy |
|---|---|---|
| [16] Chinese license plates | — | 98.5% |
| [17] LPRNet | — | 95 % |
| [18] Bangladesh license plates 1 | — | 99.6% |
| [11] Bangladesh license plates 2 | 90.9% | 98.2% |
| [19] Efficient NN | 90.93% | — |
| [20] UK number plate | — | 98.9% |
| [21] Greece LPR | 95 % | — |
| [22] Practical LPR | 95.6% | — |
| [23] UK binary character | 97.3% | — |

**Conclusion**

The research paper aims at general familiarization with the OCR and LPR systems based on all collected information. It considers the basic principles of OCR work, the works of other researchers and their proposed methods.

This article is one of the most important stages preceding presentation of the dissertation paper focusing on the development and comparison of OCR models using neural networks.

REFERENCES

1. Wikipedia free, multilingual online encyclopedia [Electronic resource] URL: https://ru.qaz.wiki/wiki/Automatic_number-plate_recognition (accessed 10.04.2021)
2. Pansare, S., & Joshi, D. (2014). A Survey on Optical Character Recognition Techniques. *International Journal of Science and Research (IJSR)*, *3*(12), 1247-1249.
3. Singh, S. (2013). Optical character recognition techniques: a survey. *Journal of emerging Trends in Computing and information Sciences*, *4*(6), 545-550.

4. Rao, N. V., Sastry, A. S. C. S., Chakravarthy, A. S. N., & Kalyanchakravarthi, P. (2016). OPTICAL CHARACTER RECOGNITION TECHNIQUE ALGORITHMS. *Journal of Theoretical & Applied Information Technology*, *83*(2).
5. Islam, N., Islam, Z., & Noor, N. (2017). A survey on optical character recognition system. *arXiv preprint arXiv:1710.05703*.
6. Phangtriastu, M. R., Harefa, J., & Tanoto, D. F. (2017). Comparison between neural network and support vector machine in optical character recognition. *Procedia computer science*, *116*, 351-357.
7. Tarigan, J., Diedan, R., & Suryana, Y. (2017). Plate recognition using backpropagation neural network and genetic algorithm. *Procedia Computer Science*, *116*, 365-372.
8. Saif, A. S., Hossain, M. J., Uzzaman, M. H., Rahman, M., & Islam, M. T. (2018). Real Time Bangla Vehicle Plate Recognition towards the Need of Efficient Model-A Comprehensive Study. *International Journal of Image, Graphics and Signal Processing*, *10*(12), 29.
9. Patel, C., Shah, D., & Patel, A. (2013). Automatic number plate recognition system (anpr): A survey. *International Journal of Computer Applications*, *69*(9).
10. Sahu, N., & Sonkusare, M. (2017). A study on optical character recognition techniques. The International Journal of Computational Science, Information Technology and Control Engineering, 4, 1-14.
11. Ahsan, M., Based, M., & Haider, J. (2021). Intelligent System for Vehicles Number Plate Detection and Recognition Using Convolutional Neural Networks. *Technologies*, *9*(1), 9.
12. Silva, S. M., & Jung, C. R. (2018). License plate detection and recognition in unconstrained scenarios. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 580-596).
13. Asthana, S., Sharma, N., & Singh, R. (2011). Vehicle number plate recognition using multiple layer back propagation neural networks. *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, *1*(1).
14. Sulehria, H. K., Zhang, Y., Irfan, D., & Sulehria, A. K. (2008). Vehicle number plate recognition using mathematical morphology and neural networks. *Wseas transactions on computers*, *7*(6), 781-790.
15. Gilly, D., & Raimond, K. (2013). A survey on license plate recognition systems. *International Journal of Computer Applications*, *61*(6).
16. Xu, Z., Yang, W., Meng, A., Lu, N., Huang, H., Ying, C., & Huang, L. (2018). Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 255-271).
17. Zherzdev, S., & Gruzdev, A. (2018). Lprnet: License plate recognition via deep neural networks. *arXiv preprint arXiv:1806.10447*.
18. Dhar, P., Guha, S., Biswas, T., & Abedin, M. Z. (2018, February). A system design for license plate recognition by using edge detection and convolution neural network. In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.
19. Caner, H., Gecim, H. S., & Alkar, A. Z. (2008). Efficient embedded neural-network-based license plate recognition system. *IEEE Transactions on Vehicular Technology*, *57*(5), 2675-2683.
20. Rhead, M., Gurney, R., Ramalingam, S., & Cohen, N. (2012, October). Accuracy of automatic number plate recognition (ANPR) and real world UK number plate problems. In *2012 IEEE International Carnahan Conference on Security Technology (ICCST)* (pp. 286-291). IEEE.
21. Anagnostopoulos, C. N. E., Anagnostopoulos, I. E., Psoroulas, I. D., Loumos, V., & Kayafas, E. (2008). License plate recognition from still images and video sequences: A survey. *IEEE Transactions on intelligent transportation systems*, *9*(3), 377-391.

22. Oz, C., & Ercal, F. (2005, June). A practical license plate recognition system for real-time environments. In *International Work-Conference on Artificial Neural Networks* (pp. 881-888). Springer, Berlin, Heidelberg.
23. Zhai, X., Bensaali, F., & Sotudeh, R. (2012, July). OCR-based neural network for ANPR. In *2012 IEEE International Conference on Imaging Systems and Techniques Proceedings* (pp. 393-397). IEEE.

**Салерова Д.К., Сарсембаев А.А.**
**Таңбаларды оптикалық тануды пайдалану арқылы нөмірлер белгілерін тануға шолу мақаласы**

**Аңдатпа.** Технологияның қарыштап дамуы халық санының артуымен нөмірлерді тану жүйелеріне деген қажеттілік – ең өзекті мәселе. Полицейлер, саябақ басқармасы, жол қозғалысын басқару өз жұмысын жақсарту үшін LPR пайдаланады. LPR жүйесі сандық фотокамерадан кескін түсіреді, содан кейін алдын ала өңдеуді, таңбаларды сегментациялауды және таңбаларды тануды жүзеге асырады. Бұл мақала OCR әдістері мен қадамдарын, LPR қосымшаларын көрсететін болып табылады және басқа да зерттеушілер қолданатын алгоритмдер мен әдістерге қысқаша шолу жасалып, сипаттама береді.

**Түйінді сөздер**: нөмірлерді тану жүйелер (LPR), таңбаларды оптикалық тану (OCR), OCR алгоритмдер.

**Салерова Д.К., Сарсембаев А.А.**
**Обзорная статья распознавания номерных знаков с использованием оптического распознавания символов**

**Аннотация.** С развитием технологий и увеличением населения потребность в системах распознавания номерных знаков становится все более и более актуальной. Полиция, управление парками, управление дорожным движением используют LPR для улучшения своей работы. Система LPR захватывает изображение с цифровой камеры, затем выполняет предварительную обработку, сегментацию и распознавание символов. Данная статья представляет собой обзор, который показывает методы и этапы OCR, области применения LPR и дает краткое описание алгоритмов и методов, применяемых другими исследователями.

**Ключевые слова:** оптическое распознавание номерных знаков, оптическое распознавание символов, алгоритмы оптического распознавания, частота ошибок символов.

**Авторлар туралы мәлімет:**

**Салерова Диана Кайратовна**, бакалавр, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының магистрлік студенті, Халықаралық ақпараттық технологиялар университеті.

**Сарсембаев Айдос Айдарович**, PhD, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының ассистент-профессоры, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Салерова Диана Кайратовна**, бакалавр, магистрант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**Сарсембаев Айдос Айдарович**, PhD, ассистент-профессор кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**About the authors:**

**Diana K. Salerova**, master student, Department of Computer Engineering and Information Security, International Information Technology University.

**Aidos A. Sarsembayev**, PhD, Assistant-Professor, Department of Computer Engineering and Information Security, International Information Technology University.

УДК 004.932.72'1

**Salerova D.K. [1], Sarsembayev A.A.[2]**

[1,2] International Information Technology University, Almaty, Kazakhstan

# RESEARCH ON THE EXISTING IMAGE CLASSIFICATION METHODS

**Abstract.** Image classification is a topical issue in data science, which is applied in many areas. It is the process of getting classes of information from a multichannel bitmap using specific rules. There are three approaches to image classification - supervised, unsupervised and object-based. Support Vector Machines, Artificial Neural Network, Decision Tree, Convolutional Neural Network are the solution methods used for image classification. Choice of the classification approach depends on the way an analyst interacts with the computer. The whole process boils down to image stack acquisition, preprocessing and classification.

**Keywords:** image processing, Support Vector Machines, Artificial Neural Network, Decision Tree, Convolutional Neural Network.

## Introduction

The problem of image classification consists in getting the image as an input and outputting its class or a group of probable classes that best characterize the image. For people, this is completely natural and simple. It is one of the first skills from birth. On the contrary, the computer "sees" only the pixels of the image, which have different colors and intensity. The classification process falls down into the following steps: preprocessing of digital data, feature extraction, selection of training data, decision and classification.

Supervised classification uses spectral signatures obtained from training samples. It means that the user can select sample pixels in an image and use them as references for the classification of all other pixels.

Unsupervised classification finds spectral classes (or clusters) in the multi-image without the analyst’s intervention. The computer uses special techniques to determine the related pixels and group them into classes.

Object-based classification segments an image by grouping pixels based on their spectral characteristics and generates objects with different geometries.

At the same time, it is hard to determine and classify an object, because of the variation between the images of same class, viewpoints, scales or background clutter.

Image classification solves various problems in such areas as medicine, environmental change, education, agriculture, object detection and security.

## Related work

The paper by Sunayana G. Domadia [1] describes supervised and unsupervised image classification techniques by implementing and analyzing their accuracy and time. The author uses the k-means algorithm based on a minimum distance while other algorithms are based on probability distribution.

Desheng Liu et al. [2] explain the advantages and the limitations of an object-based approach in image classification, in comparison with a pixel-based approach. Their conclusion is based on the differences in the classification functions and classification units.

The paper by Abass Olaode [3] identifies dimension reduction and clustering algorithms as the main classes needed in unsupervised image classification.

Le Hoang Thai [4] describes image classification in terms of the process of implementing Artificial Neural Network together with Support Vector Machine.

Farhana Sultana [5] explains the use of different Convolutional Neural Network architectures for image classification.

The paper by Majid Shadman Roodposhti [6] shows the assessment techniques which do not depend on the test data. It compares the efficiency of using deep neural networks (DNN) with the well-known random forest (RF) technique.

**Image classification process**

The stages of image classification are digital data, preprocessing, feature extraction, selection of training data, decision and classification, classification output and accuracy assessment. As demonstrated in Figure 1, the first stage is about downloading image files from the web. Then damaged and non-image files, small images should be removed, large images should be downscaled, and the remaining part should be encoded in RGB. The third stage is the determination or computation or finding out the characteristics from the image sample. The next stage describes the choice of the specific attribute that best characterizes the pattern. Then in the decision and classification stage the detected objects are classified into predetermined classes by applying an appropriate method. This method compares image sample with the target one. The classification output is the prediction result with a suitable label attached to the object. The last stage, called accuracy assessment, describes the recognition of probable sources of errors. Also, it is used in the comparison as a detector.

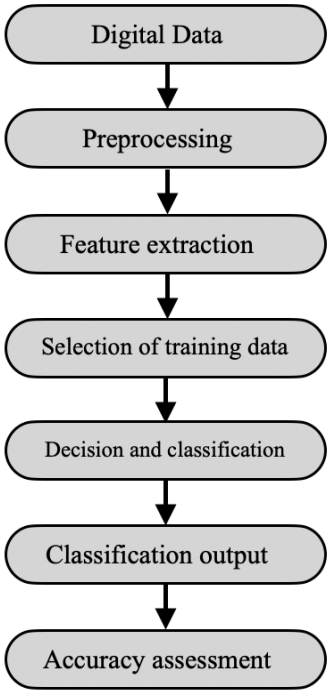



*Figure 1 - Image classification steps*

**Application**

Image classification is applied in different areas such as:

- Automatic control in production applications;
- Assisting humans with identification tasks such as a species identification system;
- Control processes for industrial robots;
- Event detection for visual observation or people counting;
- Interaction tasks for human-computer interaction devices;
- Modeling objects or environments, such as medical image analysis or topographic modeling;

• Navigation by an autonomous vehicle or a mobile robot;
• Organizing information for indexing image databases and image sequences.

**Most popular classification methods**

Convolutional Neural Network (CNN) is the most famous model applied in image classification. The main advantage of CNN is that the model understands the image on the local level. A small number of parameters significantly reduces the time and the amount of data needed to train a model.

Support Vector Machines (SVM) are the supervised learning models, where the algorithm builds hyperplanes to divide the data into classes. SVM is accurate in high dimensional spaces and uses memory effectively because of a subset of training points in the decision function.

Artificial Neural Network (ANN) is a computing system that comes from biological neural networks. ANN have adaptive weights on the way between neurons, that can be regulated by a learning algorithm. To improve the model, this algorithm learns from the observed data.

Decision Trees (DT) is a supervised learning technique, that can train samples and create a decision tree. DT consists of levels called root nodes, sub-nodes and leaves. The nodes are connected by branches that show the classification direction.

**Benefits and drawbacks**

Table 1 below presents a comparison of the benefits and drawbacks of various image classification methods.

Table 1 - Classification methods comparison

| | Advantages | Disadvantages |
|---|---|---|
| Convolutional Neural Network (CNN) | automatically detects the important features, computationally efficient, highly accurate | high computational cost |
| Support Vector Machines (SVM) | effective in high dimensional spaces, memory efficient, less risk of overfitting | not suitable for and takes a long training time with large data sets, depends on noise |
| Artificial Neural Network (ANN) | parallel processing capability, a distributed memory, very efficient for large data sets, robust to noise in the training data | high computational cost |
| Decision Trees (DT) | requires less effort for data preparation during pre-processing, does not require normalization of data | Takes longer time to train a model, a small change in data can cause a large change in the structure |

**Conclusion**

In this paper we have analyzed the supervised, unsupervised and object-based approaches to image classification. We have defined the most popular classification methods and have figured out their benefits and drawbacks. This will help researchers choose the most suitable classification method according to their requirements.

Convolutional Neural Networks and Artificial Neural Networks are most suitable and popular methods for solving image classification problems. One of their common disadvantages is high computational cost, but the benefits of using them are much greater.

REFERENCES

1. Domadia, S. G., & Zaveri, T. (2011, May). Comparative analysis of unsupervised and supervised image classification techniques. In *Proceeding of National Conference on Recent Trends in Engineering & Technology* (pp. 1-5).
2. Liu, D., & Xia, F. (2010). Assessing object-based classification: advantages and limitations. *Remote Sensing Letters*, *1*(4), 187-194.
3. Olaode, A., Naghdy, G., & Todd, C. (2014). Unsupervised classification of images: A review. *International Journal of Image Processing*, *8*(5), 325-342.
4. Thai, L. H., Hai, T. S., & Thuy, N. T. (2012). Image classification using support vector machine and artificial neural network. *International Journal of Information Technology and Computer Science*, *4*(5), 32-38.
5. Sultana, F., Sufian, A., & Dutta, P. (2018, November). Advancements in image classification using convolutional neural network. In *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)* (pp. 122-129). IEEE.
6. Shadman Roodposhti, M., Aryal, J., Lucieer, A., & Bryan, B. A. (2019). Uncertainty assessment of hyperspectral image classification: Deep learning vs. random forest. *Entropy*, *21*(1), 78.
7. Anthony, G., Greg, H., & Tshilidzi, M. (2007). Classification of images using support vector machines. *arXiv preprint arXiv:0709.3967*.

**Салерова Д.К., Сарсембаев А.А.**

**Қолданыстағы бейнелерді жіктеу әдістерін зерттеу**

**Аңдатпа.** Кескіндерді жіктеу – көптеген салаларда қолданылатын деректанудағы өзекті тақырып. Бұл нақты ережелерді қолдана отырып, көп жолақты растрлық картадан ақпарат алу процесі. Кескінді жіктеудің үш әдісі бар: бақыланатын оқыту, бақылаусыз оқыту және объектілік. Олардың қолданылуы классификация кезінде талдаушының компьютермен өзара әрекеттесуіне байланысты болады. Процестің барлығы кескіндер стегін алудан, алдын ала өңдеу мен жіктелуден тұрады. Тірек векторлық машиналар, жасанды нейрондық желілер, шешім ағаштары, конволюциялық нейрондық желілер – кескіндерді жіктеу әдістері.

**Түйінді сөздер**: кескінді өңдеу, тірек векторлық машиналар, жасанды нейрондық желі, шешім ағашы, конволюциялық нейрондық желі.

**Салерова Д.К., Сарсембаев А.А.**

**Исследование существующих методов классификации изображений**

**Аннотация.** Классификация изображений является актуальной темой в науке о данных, которая применяется во многих областях. Это процесс получения классов информации из многоканального растрового изображения с использованием определенных правил. Существует три метода классификации изображений: обучение с учителем, обучение без учителя и объектный. Их применение зависит от того, как аналитик взаимодействует с компьютером во время классификации. Весь процесс состоит из получения стека изображений, предварительной обработки и классификации. Машины опорных векторов, искусственная нейронная сеть, дерево решений, сверточная нейронная сеть — все это методы классификации изображений.

**Ключевые слова:** обработка изображений, машины опорных векторов, искусственная нейронная сеть, дерево решений, сверточная нейронная сеть.

**Авторлар туралы мәлімет:**

**Салерова Диана Кайратовна**, бакалавр, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының магистрлік студенті, Халықаралық ақпараттық технологиялар университеті.

**Сарсембаев Айдос Айдарович**, PhD, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының ассистент-профессоры, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Салерова Диана Кайратовна**, бакалавр, магистрант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**Сарсембаев Айдос Айдарович**, PhD, ассистент-профессор кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**About the authors:**

**Diana K. Salerova**, master student, Department of Computer Engineering and Information Security, International Information Technology University.

**Aidos A. Sarsembayev**, PhD, Assistant Professor, Department of Computer Engineering and Information Security, International Information Technology University.

УДК 004.051

**Orazalin A.[1], Mursaliyev D.E.[2], Sergazina A.S.[3]**

[1,2] International Information Technology University,
[3] Kazakh-British Technical University Almaty, Kazakhstan

# CURRENT CONVOLUTIONAL NEURAL NETWORK ARCHITECTURES FOR DIAGNOSING MEDICAL IMAGES

**Abstract.** This work shows the current architectures of convolutional neural networks for diagnosing medical images in the lungs and brain, the algorithms are implemented in the Python programming language using the libraries for working with neural networks. It presents the results of comparing time and resources needed for model training which demonstrate a higher accuracy of early diagnosis achieved by using the current architectures of convolutional neural networks.

**Keywords:** convolutional neural networks, medical images, pneumonia, Python, object detection.

## Introduction

The main purpose of this work is to identify relevant architectures for the possibility of early diagnosis of pulmonary diseases and building fast, low-cost models in a modern programming language.

## About CNN in medicine

Neural networks are nonlinear systems that allow of much better classification of data than the commonly used linear methods. When applied in medical diagnostics, they make it possible to significantly increase the accuracy of the method without reducing its sensitivity.

Diagnosis is a special case of classification of events, highly valuable in relation to those events that are not present in the training of a neural network set. The advantage of neural network technologies is manifested here - they are able to carry out such a classification, summarizing the previous experience and applying it to new cases.

Machine learning techniques help efficiently analyze medical images, namely, when identifying tumors in CT scans, magnetic resonance imaging (MRI) and positron emission tomography (PET).

The advantages of automatic processing of medical images are obvious:

1. Faster diagnosis of diseases.
2. Convenience of using software tools.
3. Reduced data error rate.

## About data sets

Pneumonia is one of the most frequent causes of mortality among older persons and children.

According to WHO, in 2017, the cause of death from this disease was globally diagnosed in 15% of cases among children up to 5 years. Diagnosing pneumonia is challenging, as beyond radiographic analysis such a diagnosis requires confirmation by the medical history data. In X-ray images pneumonia usually manifests itself as a region or regions of increased opacity [6]. However, its diagnosis is difficult due to a number of other conditions in lungs such as fluid (pulmonary edema), bleeding, volume loss (atelectasis or collapse), lung cancer, post-radiation or surgical changes.

In order to help radiologists diagnose pneumonia, North American Radiological Association (RSNA) in collaboration with the US National Institutes of Health, Thoracic Society radiology and MD.ai organized a discovery competition on diagnosing pneumonia from X-rays on the Kaggle

platform. The supplied dataset contained 26,684 unique X-ray images, containing 3 classes of labels (without pathology - 29%, with pathology, but without darkening - 40%, with darkening - 31%). All images with blackouts were marked as rectangular areas indicative of pneumonia. The images were divided into training (23,115), validation (2569) and test (1000) samples. To reduce the impact of retraining the obtained models on a test sample at the first stage, at the end of the competition, the organizers added a closed test part containing an additional 3000 images.

**RetinaNet**

The architecture of the convolutional neural network (SNS) RetinaNet consists of 4 main parts, each of which has its own purpose:

1. Backbone is a core network used to extract features from an input image. This part of the network is variable and its basis may include classification neural networks such as ResNet, VGG, EfficientNet and others;
2. Feature Pyramid Net (FPN) - a convoluted neural network built in the form of a pyramid, serving to combine the merits of maps of features of the lower and upper levels of the network, in which the former have a high resolution, but low semantic, generalizing ability; and the latter - vice versa;
3. Classification Subnet - a subnet that extracts information about object classes from the FPN, solving the classification problem;
4. Regression Subnet - a subnet that extracts information about the coordinates of objects in the image from the FPN, solving the regression problem.

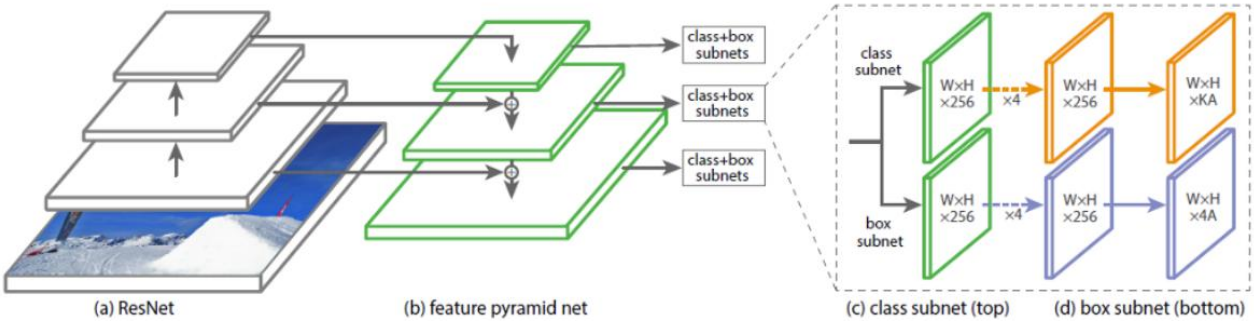



*Figure 1 - Architecture of RetinaNet s backbone-network ResNet*

**Backbone part of the network RetinaNet**

Given that the part of the RetinaNet architecture that receives the image and highlights important features is variable and the information extracted from this part will be processed in the next stages, it is important to select the appropriate backbone network for the best results.

**Mask R-CNN as a backbone for RetinaNet**

Mask R-CNN develops the Faster R-CNN architecture by adding another branch that predicts the position of the mask covering the found object, and thus solves the instance segmentation problem. The mask is simply a rectangular matrix, in which 1 at some position means that the corresponding pixel belongs to an object of a given class, 0 - that the pixel does not belong to such an object.

This ensemble of two architectures diagnoses pneumonia with a high accuracy.

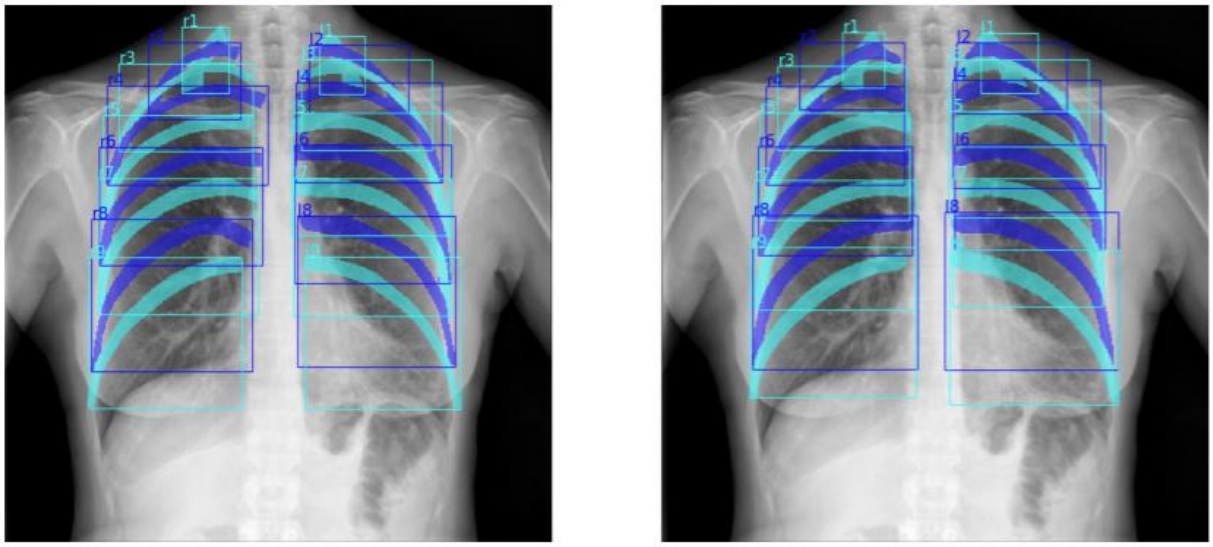


*Figure 2 - Mask R-CNN results in chest-x-ray pictures*

**VGG16**

VGG16 is a convolutional network for image extraction which is one of the most famous models sent to the ILSVRC-2014 competition. It is an improved version of AlexNet, which replaces the large filters (size 11 and 5 in the first and second convolutional layers, respectively) with several 3x3 filters one after the other.

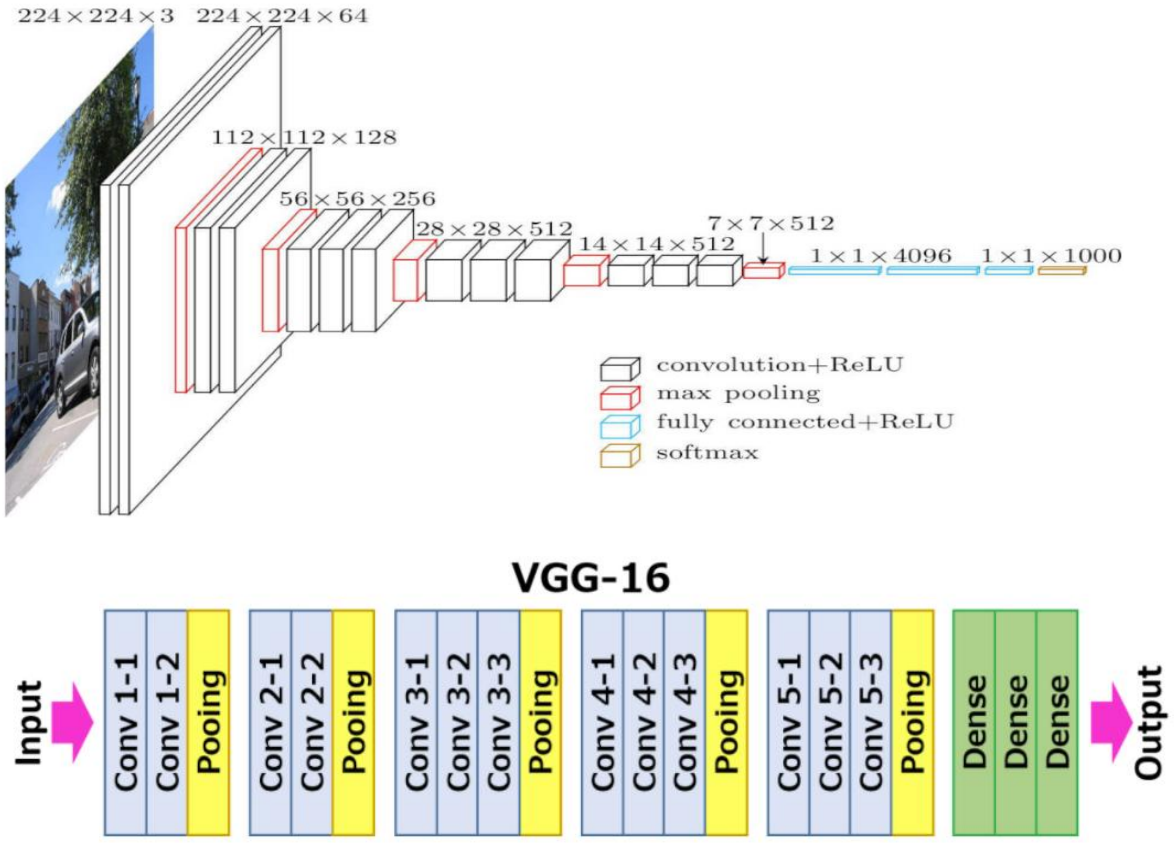



*Figure 3 - Architecture of VGG16*

The input of the conv1 layer is RGB images with the size of 224x224. The images then go through a stack of convolutional layers that use filters with a very small 3x3 receptive field (which is the smallest size for getting an idea of where the right / left, top / bottom, center is).

One of the configurations uses a 1x1 convolutional filter, which can be represented as a linear transformation of the input channels (followed by non-linearity). The convolutional step is fixed at 1 pixel. The spatial padding of the input of the convolutional layer is chosen so that the spatial resolution is preserved after convolution, that is, the padding is 1 for 3x3 convolutional layers. Spatial pooling is done using five max-pooling layers that follow one of the convolutional layers (not all convolutional layers have subsequent max-pooling layers). The max-pooling operation is performed in a 2x2 pixel window with a step of 2.

After the stack of convolutional layers (which has different depths in different architectures) there are three fully connected layers: the first two have 4096 channels, the third has 1000 channels (since the ILSVRC competition requires to classify objects into 1000 categories; therefore, one channel corresponds to a class). The last is the soft-max layer. The configuration of fully connected layers is the same in all neural networks.

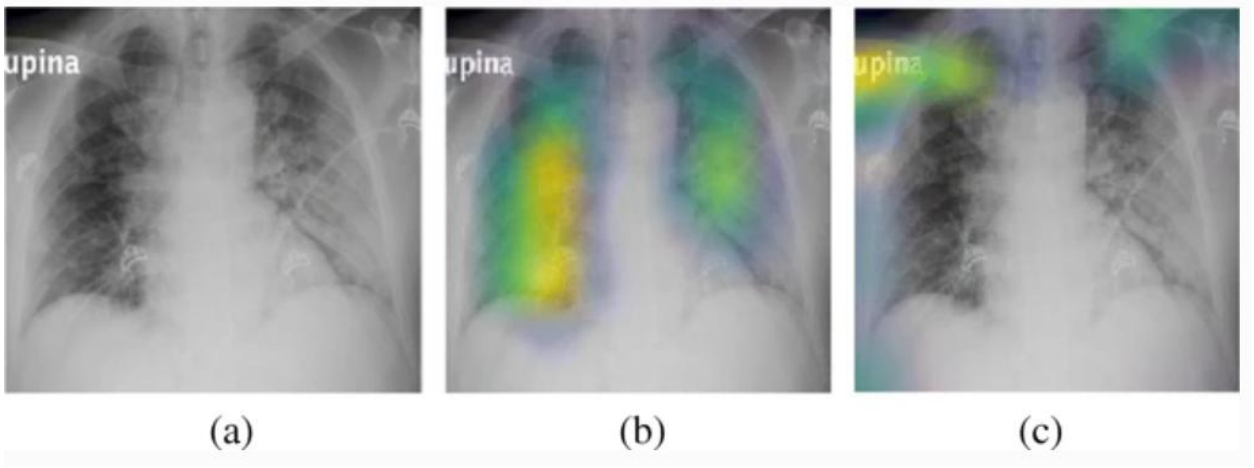



*Figure 4 - VHH16 results in chest-x-ray pictures*

Unfortunately, the VGG network has two major drawbacks:

1. Very slow learning speed.
2. The network architecture itself is too heavy (disk and bandwidth issues appear)

Due to the depth and number of fully connected nodes, VGG16 weighs over 533 MB. This makes the VGG deployment process a tedious task. Although VGG16 is used to solve many classification problems with neural networks, smaller architectures are preferred (SqueezeNet, GoogLeNet, and others). Despite its drawbacks, this architecture is a great building block for learning, as it is easy to implement.

**U-Net**

The U-Net convolutional network was designed with medical imaging in mind. It achieves high accuracy and uses small datasets.

The network is trained end-to-end on a small number of images and outperforms the previous best method (sliding window convolutional network) in the ISBI competition for segmentation of neural structures in electron microscopic stacks.

Using the same network that was trained on transmission light microscopy images, U-Net ranked first by a large margin in the 2015 ISBI competition for cell tracking in these categories.

U-Net conforms to the codec architecture. The encoder is gradually decreasing

spatial dimension by combining layers, and the decoder gradually restores the object detail and spatial dimension. There are also fast connections from encoder to decoder to help the decoder better reconstruct the object details.

The network does not have fully connected layers and only uses the valid part of each convolution, that is, the segmentation map contains only the pixels for which the full context is available in the input image. For high-quality segmentation Unet increases the amount of data by deforming the existing images.

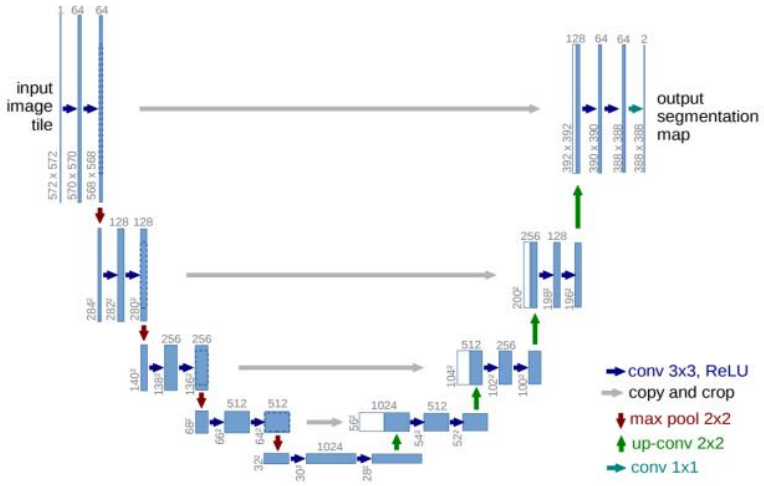



*Figure 5 - U-Net architecture*

Each blue square corresponds to a multichannel feature map. The number of channels is shown at the top of the rectangle. The x-y dimensions are shown in the bottom of the left edge of the rectangle. White rectangles represent copies of the map properties.

The U-Net consists of a declining part (left side) and an expanding part (right side).

The tapering part corresponds to a typical convolutional network architecture. It consists of the repeated use of two 3x3 convolutions (convolution without indentation), followed by the ReLU activation function. Then comes a downsampling layer with a 2x2 filter and step 2 to compact the feature map. At each step of downsampling, we double the number of feature channels.


*Figure 6 - Lung segmentation using U-Net before training the convolutional neural network: (a) the original chest X-ray image, (b) a mask of lung structures segmented through U-Net, and (c) the final segmented image of the lungs.*

REFERENCES

1. Deep Convolutional Neural Networks for Chest Diseases Detection/ Rahib H. Abiyev and Mohammad Khaleel Sallam Ma'aitah, J Health Eng, 2018, p. 5-6
2. On the experience of using artificial intelligence technologies for automatic recognition of X-ray images of the chest cavity organs, Tolkachev A. and Kuleev R, Digital Healthcare, 2015, p. 3
3. Sequential Rib Labeling and Segmentation in Chest X-Ray using Mask R-CNN, J¨oran Wessel, arXivLabs, 2019, p. 3
4. VGG16 is a convolutional network for image extraction, Milyutin Ilya, [Electronic resource] URL: https://neurohive.io/ru/vidy-nejrosetej/vgg16-model/ / (accessed: 23.11.2018)
5. RetinaNet Neural networks architecture // Konstantin (datist), [Electronic resource] URL: https://habr.com/ru/post/510560/ (accessed: 11.07.2020)
6. U-NET FOR SOLVING THE PROBLEM OF SEGMENTATION OF MEDICAL IMAGES, Kozlova O, Fifth International Scientific and Practical Conference “BIG DATA and Advanced Analytics, 2019, p. 254.

**Оразалин А., Мурсалиев Д.Е., Сергазина А.С.**
**Актуальные сверточные архитектуры нейронной сети для диагностики медицинских изображений**

**Аннотация.** В данной работе показаны актуальные архитектуры сверточных нейронных сетей для диагностирования медицинских изображений в области легких и головного мозга. Алгоритмы реализованы на языке программирования Python и при помощи полезных библиотек для работы с нейронными сетями. Представлены результаты сравнения времени и ресурсов для обучения моделей. По результатам был сделан вывод о более высокой точности раннего диагностирования с помощью актуальных на сегодняшний день архитектур сверточных нейронных сетей.

**Ключевые слова:** сверточные нейронные сети, медицинские изображения, пневмония, Python, обнаружение объекта.

**Оразалин А., Мурсалиев Д.Е., Сергазина А.С.**
**Медициналық кейіндік диагностикаға арналған конволюциялық жүйкелік желі архитектурасы**

**Аңдатпа.** Бұл жұмыста өкпе мен ми аймағындағы медициналық кескіндерді диагностикалауға арналған конволюциялық жүйке желілерінің қазіргі архитектурасы көрсетілген, алгоритмдер Python бағдарламалау тілінде және жүйке желілерімен жұмыс істеу үшін пайдалы кітапханаларды қолдана отырып жүзеге асырылады. Оқу үлгілері үшін уақыт пен ресурстарды салыстыру нәтижелері келтірілген. Нәтижелер бойынша конволюциялық жүйке желілерінің қазіргі архитектурасын қолдана отырып, ерте диагностиканың жоғары дәлдігі туралы қорытынды жасалды.

**Түйінді сөздер:** Конволюциялық жүйке желілері, медициналық бейнелеу, пневмония, Python, объектіні анықтау.

**Авторлар туралы мәлімет:**

**Оразалин Азимхан** «Математикалық және компьютерлік модельдеу» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті,.

**Мурсалиев Дәурен Ерахметұлы**, «Математикалық және компьютерлік модельдеу» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті.

**Сергазина Айым Серикбекқызы**, «Ақпараттық жүйелер» кафедрасының магистранты, Қазақстан-Британ техникалық университеті.

**Сведения об авторах:**

**Оразалин Азимхан,** магистрант кафедры «Математическое и компьютерное моделирование», Международный университет информационных технологий.

**Мурсалиев Даурен Ерахметович,** магистрант кафедры «Математическое и компьютерное моделирование», Международный университет информационных технологий.

**Сергазина Айым Серикбековна,** магистрант кафедры «Информационные системы» Казахстанско-Британский технический университет.

**About the authors:**

**Orazalin Azimkhan,** master student, Department of Mathematical and Computer Modeling, International Information Technology University.

**Dauren E. Mursaliyev,** master student, Department of Mathematical and Computer Modeling, International Information Technology University.

**Aiym S. Sergazina**, master student, Department of Information Systems, Kazakh-British Technical University.

УДК 004.62

**Alimkhan A.M.**

International Information Technology University, Almaty, Kazakhstan

# PREDICTING BASKETBALL RESULTS USING DEEP LEARNING ALGORITHMS

**Abstract.** With the development of information technology and an ever-expanding statistical base, the possibilities for forecasting are expanding, and the dependencies of the calculated indicators on the result are considered. In this article we compare 3 most widely spread game result prediction models, and namely, Support Vector Regression, K-Nearest Neighbor model and the Linear Regression model in terms of their prediction accuracy and experimentally demonstrate the advantages of the latter.

**Keywords:** deep learning, machine learning, prediction models, regression technique, data analysis.

## Introduction

Today, there are many new basketball metrics used by fans and analysts around the world to compare and measure players. For example, there is a Player Performance Rating (PER) that considers achievements such as field goals, free throws, 3 pointers, assists, rebounds, blocks and interceptions, and negative results such as missed shots, turns and personal fouls. Some of these advanced basketball metrics, such as PER, are great for measuring a player's performance in a basketball game. Yet, does this mean that the player will bring success to his team? May be. But the only way to see if a player succeeds is his victory. Many sports betting fans are interested in predicting the results. It is also the subject of sports analytics. Since for the most reliable forecast it is necessary to consider and correctly analyze a large complex of sports characteristics, the use of various AI methods is quite logical.

When developing your own software tool, where you can consider all the nuances of the problem like the prediction result of the game being solved, you should adjust it for yourself. However, it is obvious that this path requires much more effort and time. But it is also obvious that the result obtained will be more accurate. Naturally, to go down this path, you need not only to have sufficient qualifications in the field of artificial intelligence, but also to program well yourself. Also, it is necessary to know about deep learning, machine learning, the prediction models like Linear Regression, Support Vector Machine, and k-Nearest Neighbors Regression Model etc., to use them for predicting the result of a basketball game.

## Data analysis and exploration

Analysis of the source data. Data Preparation is a very laborious iterative process that takes up to 80% of all resource and time costs in the Data Mining life cycle and includes the following tasks for processing the initial ("raw") data [12]:

- Data examining - choice of highlights (or indicators) and objects, considering their significance for the purposes of Information Mining, quality, and specialized imperatives (volume and sort).
- Data cleansing - expelling typos, erroneous values (for illustration, a number in a string parameter, etc.), lost values (Lost Values or NA), barring copies and diverse depictions of the same protest, reestablishing uniqueness, judgment, and coherent associations.
- Feature Generation - creation of the determined highlights and their change into vectors for the Machine Learning show, as well as introduction of changes to enhance the precision of machine learning calculations.

- Integration - consolidating information from different sources (data frameworks, tables, conventions, etc.), counting their conglomeration, when unused values are calculated by summing data from a set of existing records.
- Formatting - syntactic changes that do not alter the meaning of the information, but are required for modeling devices, for sorting cases in a particular arrangement or expelling superfluous accentuation in content areas, trimming "long" words, adjusting genuine numbers to integrability, etc.

**Model Selection and Testing**

To improve the accuracy of predictive models, it is necessary to apply appropriate machine learning methods. Depending on the nature of the business problems under consideration, different approaches are used, considering the types and volumes of data. This section introduces the categories of machine learning. There are such algorithm models used in building a prediction program as:

- Linear Regression
- Support Vector Regression
- K-Nearest Neighbour

Linear Regression. Linear Regression is meant to predict continuous numeric variables. In expansion, the adjusted adaptation of straight relapse actualized within the Deductor analytical platform also allows of solving the classification problem.

Linear Regression is maybe one of the foremost well-known and popular calculations and insights in machine learning.

Predictive modeling is essentially concerned with decreasing model errors or, in other words, making predictions as accurate as possible. We will borrow calculations from different areas, counting measurements, and utilize them for this reason.

Benefits of Linear Regression:

- The quickness and ease of obtaining the model.
- Interpretability of the show. The straight show is straightforward and justifiable for the examiner. The gotten relapse coefficients can be utilized to judge how a specific calculation influences the result, and to draw extra valuable conclusions on this basis.
- Wide appropriateness. Numerous genuine forms in financial matters and trade can be depicted by direct models with adequate exactness.

Knowledge of this approach. For Linear Regression, typical problems (for example, multicollinearity) and their solutions are known, tests for assessing the static significance of the resulting models are developed and implemented.

SVM. SVM is a supervised learning algorithm used to solve classification problems. Here the most thought-of the strategy is to exchange the beginning vectors to a space of higher measurement and hunt for an isolating hyperplane with the biggest hole in this space. Two parallel hyperplanes are developed on both sides of the hyperplane isolating the classes. The isolating hyperplane will be the hyperplane that makes the most prominent remove to two parallel hyperplanes. The calculation accepts that the more noteworthy the contrast or remove between these parallel hyperplanes, the smaller the normal blunder of the classifier will be.

Advantages of SVM:

- Fast classification method.
- The method is reduced to solving a quadratic programming problem in a convex domain, which usually has a unique solution.
- The method allows for more confident classification than other linear methods.

Disadvantages of SVM:

- They have a long learning curve, so in practice they are not suitable for large datasets. Another disadvantage is that SVM classifiers do not work well with overlapping classes.

The important concepts in SVM are as follows:

- Support vectors - The data points that are closest to the hyperplane are called support vectors. The dividing line is defined using these data points.
- Hyperplane - It is a decision plane or space that is divided between a set of objects that have different classes.
- Margin - This can be defined as a gap between two lines at the data points of a cabinet of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margins are considered good margins, and small margins are considered bad ones.

The main purpose of SVM is to subdivide datasets into classes to find the maximum limit hyperplane (MMH), and this can be done in the following two steps:

- First, the SVM iteratively generates hyperplanes that best separate classes.
- It will then select the hyperplane that separates the classes correctly.

K-Nearest Neighbour. kNN stands for k Nearest Neighbor or k Nearest Neighbors - it is one of the simplest classification algorithms, sometimes used in regression problems. Due to its simplicity, it is a good example from which to start exploring the field of Machine Learning. This article describes an example of writing the code of such a classifier in Python, as well as visualizing the results obtained.

The allocation issue in machine learning is the issue of allotting a question to one of the predefined classes based on its formalized highlights. Each of the objects in this issue is referred to as a vector in an N-dimensional space, each measurement in which may be a depiction of one of the object's qualities. Let us say we need to classify monitors: the measurements in our parameter space will be the diagonal in inches, the aspect ratio, the maximum resolution, the presence of an HDMI interface, cost, etc. The case of text classification is somewhat more complicated, for this purpose the term-document matrix is usually applied.

**Literature Review**

The articles discuss the most interesting trends in machine learning and artificial intelligence formed at the beginning of 2018 outside of specific mathematical methods for optimization, processing, and analysis of data. There is an increasing attention of researchers to the question of methodologies, or metamodels (from the English metamodel): the principles of using, combining, and choosing specific models and methods of machine learning. A long-term progress in the development of machine learning methods has spawned not only a variety of mathematical, software, and even hardware solutions designed for predictive and generative data analysis tasks in various fields, but also encountered many difficulties and obstacles along the way.

The main difficulty that a person faces in the process of getting to know the field of machine learning is a huge number of disparate methods, each of which has its own features, area of use and benefits. However, this diversity sometimes baffles sophisticated researchers. With the development of mathematical and algorithmic methods it becomes more and more difficult to navigate well in all the nuances of the applied algorithms. Unfortunately, the methodological base lags far behind the rapid process of developing new learning algorithms, and the process of choosing a trainable model sometimes comes down to a simple search.

Despite the rapid development of machine learning in the last decade, artificial intelligence remains a very vague concept. It includes many subject areas, from time series prediction to generating plausible images on a specific topic. Machine learning methods, which form the computational basis of artificial intelligence technologies, remain highly specialized for each specific task.

Support vector machines - the concept of the algorithm is, as in the case of logistic regression, in the search for a dividing plane (or several planes), however, the way to search for this plane in this case is different - a plane is searched so that the distance from it to the nearest points - representatives of both classes – is as long as possible, for which methods of quadratic optimization are usually applied.

Lazy classifiers (Lazy learners) - a special kind of classification algorithms, which, instead of pre-building a model and then, based on it, making decisions about assigning an object to a particular class, hinge on the idea that similar objects most often have one and the same class. When such an algorithm receives an object for classification as an input, it searches among previously viewed similar objects and, using information about their classes, forms its prediction regarding the class of the target object.

In their article "Machine learning approaches to predicting basketball game outcomes" Sushma Jain and Harman deep Kaur discuss the difference between the execution of the NBAME model with conventional machine learning classifiers and show that the NBAME demonstrates the next expectation exactness rate (74.4%). To illuminate the inadequacies of SVMs, Pai et al. and Kaur et al. utilized SVMs combining separate choice rules and fluffy rules, to create unused prescient coordinate result models. The result illustrated that the models accomplished higher precision than did the routine SVMs.

In article of Chenjie Cao entitled "Sports Data Mining Technology Used in Basketball Outcome Prediction" the author focused on two machine learning models, SVM and hybrid fuzzy-SVM (HFSVM) to predict NBA games. Their dataset is restricted to the regular season 2015-2016, which is split into training and test sets. The hybrid Fuzzy-SVM model combines the advantages of the fuzzy model and the SVM for a basketball match result prediction. They claim to have been able to get better game results in terms of their accuracy, it is observed that this study, with its 79.2% accuracy rate achieved by using the suggested CNFS model, has the highest accuracy rate among all the studies.

You can see that classification algorithms can have a variety of ideas in their basis and, of course, for different types of problems they show different efficiency. So, for problems with a small number of input features, rule-based systems can be useful, if it is possible to calculate some similarity metric for the input objects quickly and conveniently using lazy classifiers. As for the problems with a very large number of parameters, which are difficult to identify or interpret, such as image or speech recognition, neural networks are becoming the most appropriate classification method.

**Experimental part**

There is a process of creating prediction models below with math statistics like the mean squared error, the mean absolute error and the variance score. So according to the metrics of each model, we will have a chance to select the best prediction model.

Figure 1 illustrates the process of creating the Linear Regression model:

```
# Create the Linear Regression model

linReg = linear_model.LinearRegression()
linReg.fit(x_train, y_train)

linReg.predict(x_test)

y_lin_pred = linReg.predict(x_test)

print('Score: %.3f' % linReg.score(x_train, y_train))
print('Mean squared error: %.3f' % mean_squared_error(y_test, y_lin_pred))
print('Mean Absolute error: %.3f' % mean_absolute_error(y_test, y_lin_pred))
print('Variance score: %.3f' % r2_score(y_test, y_lin_pred))
```

```
Score: 0.910
Mean squared error: 1.007
Mean Absolute error: 0.761
Variance score: 0.880
```

*Figure 1 - Linear Regression Model*

Figure 2 shows the creation of the Support Vector Regression model with math points

```
# Create the Support Vector Regression model

svr = SVR(kernel='rbf', gamma=1e-3, C=150, epsilon=0.3)
svr.fit(x_train, y_train.values.ravel())

y_svr_pred = svr.predict(x_test)

print('Score: %.3f' % svr.score(x_train, y_train))
print("Mean squared error: %.3f" % mean_squared_error(y_test, y_svr_pred))
print('Mean Absolute error: %.3f' % mean_absolute_error(y_test, y_svr_pred))
print('Variance score: %.3f' % r2_score(y_test, y_svr_pred))
```

```
Score: 0.936
Mean squared error: 1.562
Mean Absolute error: 1.038
Variance score: 0.814
```

*Figure 2 - Support Vector Regression Model*

Figure 3 demonstrates the creation of the k-Nearest Neighbors Regression Model model with math points.

```
# Create the k-Nearest Neighbors Regression Model

knn = neighbors.KNeighborsRegressor(n_neighbors = 7, weights = 'uniform')
knn.fit(x_train, y_train)

y_knn = knn.predict(x_test)

print('Score: %.3f' % knn.score(x_train, y_train))
print("Mean Squared Error: %.3f" % mean_squared_error(y_test, y_knn))
print('Mean Absolute error: %.3f' % mean_absolute_error(y_test, y_knn))
print('Variance Score: %.3f' % r2_score(y_test, y_knn))
```

```
Score: 0.823
Mean Squared Error: 2.132
Mean Absolute error: 1.215
Variance Score: 0.747
```

*Figure 3 - k-Nearest Neighbors Regression Model*

After testing 25% of the information (the rest of the information is being prepared for training), the chosen models showed the following results:

| | Model | Mean Squared Error | Mean Absolute Error | Variance Score |
|---|---|---|---|---|
| 0 | Linear | 1.020 | 0.761 | 0.879 |
| 1 | Support Vector | 1.562 | 1.038 | 0.814 |
| 2 | k-Nearest Neighbors | 2.132 | 1.215 | 0.747 |

*Figure 4 - Selected models*

And, definitely, the winner of these models is the Linear Regression Model which has a lower root mean square error (the lower the better), and a higher estimate of variance (the higher the better). This does not mean that the other two models - the Pivot Vector and k-Nearest Neighbors

Regression can be ignored, because they still have very impressive results, but not as strong as the Linear Regression Model.

**Prediction**

There are several datasets of 2017 and 2018 seasons. So here the prediction will be about the individual player win share.

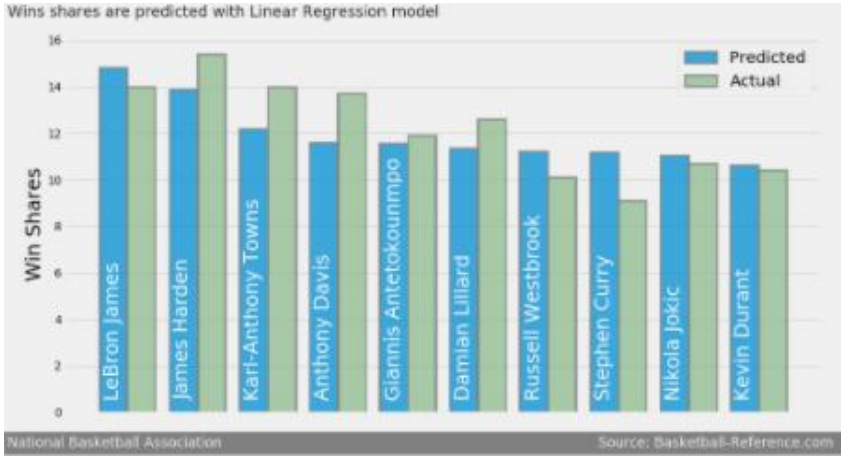



*Figure 5 - Win Share Linear Regression*

Based on the predictions of the Linear Regression Model (Figure 5), James LeBron "LABron" came out on top! The model predicted his lead in the NBA with 14.81. However, in fact, LeBron (14 winning shares) came in second after 2017-2018. MVP James Harden (15.4 winning shares) was second in the forecast. Not a bad forecast! Karl-Anthony Towns and Anthony Davis came in 3rd and 4th, respectively, in both the linear regression prediction and the last NBA season.

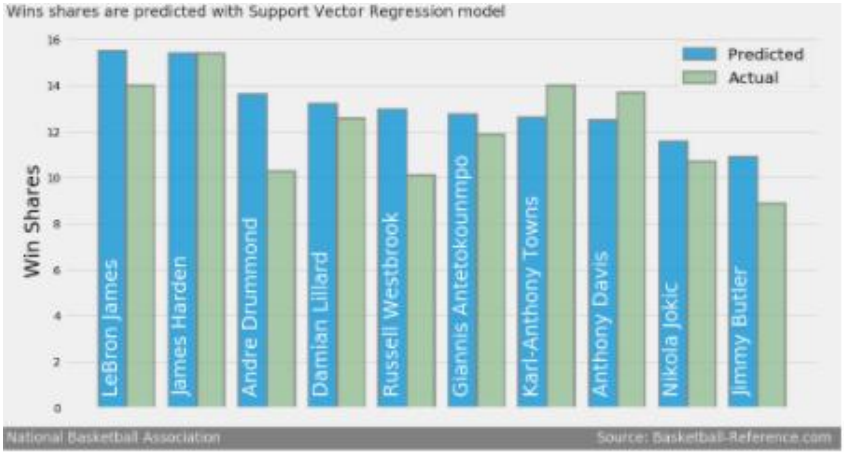



*Figure 6 - Support Vector Regression*

Win share predictions using the support vector regression model.

The Support Vector Regression Model produced unusual results which are shown in Figure 6. Based on the predictions of this model, LeBron (15.5 winning shares) came out on top again, while Harden came in second (15.4 wins). The most exciting result of this was that the model accurately predicted Harden's winning shares! He led the 2017–2018 NBA season with 15.4 winning shares, in line with the projected value. André Drummond, who was not in the top 10 players with predicted payoff using a Linear Regression model, came in third in terms of payoff using a Support Vector

Regression Model. Towns and Davis made the list of predictions, while Stephen Curry and Kevin Durant didn't even make the top ten.

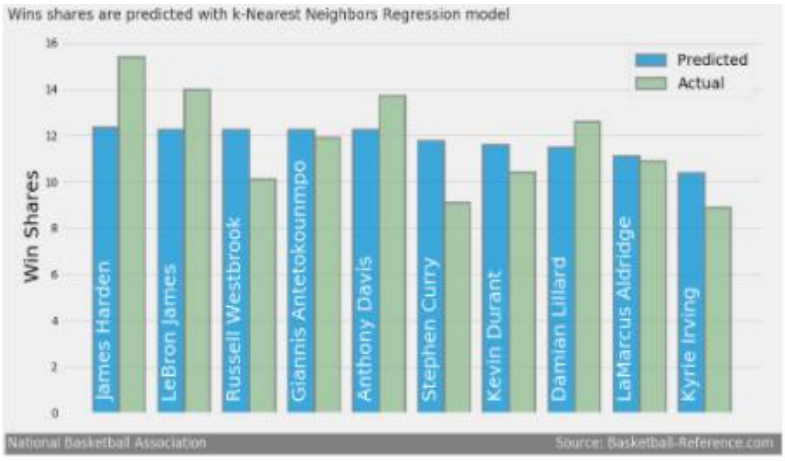



*Figure 7 - k-Nearest Neighbors Regression*

Win predictions using the k-Nearest Neighbors Regression Model

Winner share predictions using the k-Nearest Neighbors model were much less accurate than those of the previous models. For this model (Figure 7), Harden's predicted winning shares were the highest at just 12.3. Another bizarre result of the projected values was that LeBron, Russell Westbrook, Giannis Antotokumpo, and Davis were all associated with 12.24 winning shares!

**Conclusio**

In conclusion, the experimental comparison of the three prediction models has proved that the Linear Regression Model has got a high variance score, minimum scores in error tables and high calculation speed. The Linear Regression Model is also transparent and understandable for the analyst. Based on the obtained regression coefficients, one can judge how a particular factor affects the result and draw additional useful conclusions on this basis. Besides, there are many real processes in the economy and business that can be described with sufficient accuracy by Linear Regression Models for which typical problems (for example, multicollinearity) and their solutions are known, tests for assessing the static significance of the resulting models are developed and implemented.

REFERENCES

1. Regression Tree Model for Predicting Game Scores for the Golden State Warriors in the National Basketball Association [Electronic resource] URL: https://www.mdpi.com/2073-8994/12/5/835/htm (accessed: 16.05.2020)
2. Regression Techniques you should know [Electronic resource] URL: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/ (accessed: 14.09.2015)
3. Building My First Machine Learning Model | NBA Prediction Algorithm [Electronic resource] URL: https://towardsdatascience.com/building-my-first-machine-learning-model-nba-prediction-algorithm-dee5c5bc4cc1 (accessed: 9.07.2020)
4. K-Nearest Neighbors Algorithm Using Python [Electronic resource] URL: https://www.edureka.co/blog/k-nearest-neighbors-algorithm/ (accessed: 20.11.2020)
5. Implementing SVM and Kernel SVM with Python's Scikit-Learn [Electronic resource] URL: https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/ (accessed: 24.04.2020)

**Әлімхан А.М.**

**Терең оқыту алгоритмдерін қолдана отырып, баскетбол нәтижелерін болжау**

**Аңдатпа**. Бұл мақалада ақпараттық технологиялардың дамуымен және үнемі өсіп келе жатқан статистикалық базамен болжау мүмкіндіктері кеңейіп, есептелген көрсеткіштердің нәтижеге тәуелділігі ескерілетіні жазылған. Сондай-ақ нәтижені болжау барысында болжаудың әртүрлі басқа модельдерін қолдануымыз керек және болжау моделінің нәтижесінде көп ұпайлары бар модельді таңдау үшін оларды салыстырамыз.

**Түйінді сөздер***:* терең оқыту, машиналық оқыту, болжау модельдері, регрессия әдісі, деректерді талдау.

**Әлімхан А.М.**

**Прогнозирование результатов игры в баскетбол с использованием алгоритмов глубокого обучения**

**Аннотация**. В статье говорится о том, что с развитием информационных технологий и постоянно пополняющейся статистической базой расширяются возможности прогнозирования, учитываются зависимости рассчитываемых показателей от результата. Нам необходимо использовать разные модели прогнозирования, и мы сравниваем их, чтобы выбрать ту модель, которая дает возможность получить наиболее точные результаты.

**Ключевые слова:** глубокое обучение, машинное обучение, модели прогнозирования, метод регрессии, анализ данных.

**Автор туралы мәлімет:**

**Әлімхан Абайхан Мамытханұлы**, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторе:**

**Әлімхан Абайхан Мамытханұлы**, магистрант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

**About the author:**

**Alimkhan Abaikhan Mamitkhanuly**, master student, Department of Computer Engineering and Information Security, International Information Technology University.

УДК 004.8

**Адырбек Ж.А.[1], Сатыбалдиева Р.Ж.[2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан

# АНАЛИЗ ПРОЦЕССОВ ПЛАНИРОВАНИЯ И РЕШЕНИЯ ПРОБЛЕМ В ЛОГИСТИКЕ С ПОМОЩЬЮ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ

**Аннотация.** С интеллектуальной системой компания экономит время на планировании маршрутов городской дистрибуции и деньги на бензине и зарплатах. В статье представлены частичные результаты исследования, ставшие основой разработки интеллектуальной системы автоматизации транспортной логистики. Система предназначена для FMCG-сектора, интернет-магазинов, ритейлеров, дистрибьюторов и всех других секторов, где есть отдел транспортной логистики. Исследование проведено для того, чтобы понять, как устроены процессы в крупных компаниях, какие проблемы в них есть и как мы можем улучшить положение дел. Для решения проблем с помощью интеллектуальной системы был проведен опрос директоров по логистике, в ходе которого были определены специфика их работы, предпочтения и проблемы в процессе. Кроме того, была проведена работа «в поле» — был проведен обход по точкам вместе с водителями и зафиксировано время, которое им требуется для выполнения заказа. Многие компании терпят значительные финансовые издержки потому, что не могут оптимизировать рабочее время сотрудников отдела логистики. Были собраны данные, проанализированы согласно процессу, далее был построен маршрут, запущены курьеры по этим маршрутам и снова проведен анализ.

**Ключевые слова**: интеллектуальная система автоматизации транспортной логистики, автоматизация транспортной логистики, автоматизация бизнес-процессов в транспортной логистике, модель бизнес-процесса интеллектуальной системы в логистике.

## Введение

Интеллектуальная система — это техническая или специализированная система, компетентная в решении проблем, традиционно считающихся творческими, относящихся к определенному предметному диапазону, информация о которых хранится в памяти такой структуры. Структура интеллектуальной системы включает три основных блока: информационную базу, решатель и интеллектуальный интерфейс. В последние годы транспортная логистика пережила много социальных, экономических и технологических изменений. Такие преобразования значительно влияют на всю логистическую цепочку от принятия заказа до распределения заказов клиентам [1].

Актуальность работы определяется необходимостью совершенствования инструментария системы поддержки принятия решений (далее — СППР) в логистике, являющейся одним из главных направлений инвестирования, так как вложения в логистику определяют возможности поддержки бизнеса. СППР используются для определения влияния стратегических вопросов в области логистики и определения наиболее эффективных процессов, которые должны выполняться из-за стратегических вопросов с наивысшим положительным или негативным влиянием на логистику. Сегодня логистика стала неотъемлемой частью любого бизнеса. Логистикой называется многоступенчатый процесс, который управляет материальными и информационными потоками в сферах производства и обращения. Под логистикой подразумевается организация и координация процессов закупок, транспортировки продукции, оптимизация издержек всех функциональных областей предприятия, которые неизбежны на любом производстве. Логистика позволяет рассмотреть совокупность всех звеньев производственного процесса как единую систему.

Интеллектуальная система автоматизации транспортной логистики дает возможность автоматического планирования с отправкой маршрута напрямую водителю [2].

Система состоит из множества модулей, которые специфичны и настраиваются под требования клиента. Данная функция помогает в короткое время подготовить проект только с теми требованиями, которые нужны клиенту. В системе определяются виды курьеров, доставок и транспорта, в учет берутся рабочее время на точке, перерывы на обед. Исходя из этого, систему можно настроить под бизнес-процесс, который пожелает увидеть клиент.

Распределение времени по сотрудникам строится на основе алгоритмов, которые учитывают пробки, характеристики заказа (например, для перевозки замороженных продуктов требуется машина с рефрижератором) и форс-мажоры, если клиент перенес время или отменил заказ [3].

Маршрут сотрудника строится на карте, и логист может в реальном времени наблюдать за прогрессом или вносить корректировки. Кроме того у сотрудника есть возможность отмечать прогресс у себя, что позволяет передавать клиенту информацию о точном времени доставки.

Во втором разделе описывается проблема в предприятии, которая была обнаружена при изучении работы сотрудников транспортной логистики во время совместного выезда. В этом же разделе предоставляются рекомендации и решения по планированию маршрута. В третьем разделе предлагаются выводы касательно проведенного исследования.

**Автоматизация бизнес-процессов предприятия**

FMCG и другие крупные компании часто не могут спрогнозировать время доставки. Из-за этого многие заказы оказываются просроченными, что огорчает клиентов и влияет на возврат заказанных товаров. У многих предприятий уровень сервиса и качества попадания в заявленные временные окна составляет меньше 80%.

Сейчас на рынке уже существуют сервисы для оптимизации логистики: «Maxoptra», «Деливератор» и SAP. Но они либо недостаточно гибкие, либо дорогие (стоимость внедрения SAP начинается от $400 тысяч). Описываемое в данной статье решение является альтернативным продуктом для любого вида бизнеса.

Работа «в поле» помогла обнаружить несколько неочевидных вещей:

– водители часто выбирают маршрут «на глаз», даже если он не оптимален;

– дорожная ситуация постоянно меняется (например, из-за ДТП), поэтому нужна система, которая сможет предсказывать её изменения;

– необходимо регулярно обновлять и корректировать маршруты между заказами.

Цель выезда — это исследование предпроектной стадии, т. е. изучение оптимального перевода текущей работы сотрудников транспортной логистики в интеллектуальную систему.

Был совершен выезд с несколькими водителями, чтобы посмотреть, как они работают, и сравнить их реальную работу на привычном маршруте с тем, как планирует маршрут система, и выявить отклонения для дальнейшей корректировки настройки системы. Было сделано планирование согласно привычным для водителей районам. В то же время им было сказано двигаться не так, как запланирует система, а как они считают нужным, поясняя все особенности своего выбора маршрута. В итоге основным параметром, который необходимо скорректировать, оказались временные окна доставки в магазины, т. к. более 80% из них совершенно некорректно настроены и не соответствуют действительности. Временные окна являются неправильными, т. к. некорректная информация была внесена самими сотрудники компании.

Временное окно — это ограничение, которое необходимо выполнить при построении маршрута. Например, для временного окна заказа выполнение ограничения означает доставку в указанном диапазоне времени.

Видно фактическое время прибытия (зеленым) в точку рядом с настроенными временными окнами (красным) и то, какое время рекомендует система (синим). Можно заметить, что система ни разу в своих планах не отклонилась от заданных временных окон.

Система — это машина, которая работает, исходя из заданных ей параметров. Согласно параметрам, которые задавали машине, она и выводит результат. Если заданные параметры не совпадают с реальностью, люди могут оказаться не довольны результатами, которые выдает система, но система не виновата в том, что ей задают изначально неправильные параметры. В компании были заведены временные окна, которые отмечены в таблице 1.

Таблица 1 - Временные окна

| Название расположения клиента* | Временные окна операции | Фактическое время прибытия | Планируемое время прибытия |
|---|---|---|---|
| Клиент 1, адрес 1 | 08:00-12:00 | 6:30 | 11:02 |
| Клиент 2, адрес 2 | 09:00-12:00 | 8:02 | 11:39 |
| Клиент 3, адрес 3 | 09:00-12:00 | 8:05 | 11:14 |
| Клиент 4, адрес 4 | 10:00-16:00 | 8:05 | 11:23 |
| Клиент 5, адрес 5 | 09:00-13:00 | 8:23 | 12:21 |
| Клиент 6, адрес 6 | 10:00-16:00 | 8:26 | 12:37 |
| Клиент 7, адрес 7 | 00:01-23:59 | 8:28 | 12:30 |
| Клиент 8, адрес 8 | 10:00-13:00 | 8:28 | 12:51 |
| Клиент 9, адрес 9 | 09:00-13:00 | 8:35 | 12:43 |
| Клиент 10, адрес 10 | 08:00-12:00 | 8:55 | 11:32 |

В целях сохранения конфиденциальности данных компании, наименование клиентов компании и их адреса были изменены на стандартные названия.

Ниже пример рекомендуемых временных окон для вышестоящих точек (Таблица 2). Рекомендации по точкам должны давать в первую очередь водители доставки, т. к. именно они владеют наибольшей информацией по особенностям взаимоотношений с точками.

Таблица 2 - Рекомендуемые временные окна

| Название расположения клиента | Рекомендуемые временные окна операции |
|---|---|
| Клиент 1, адрес 1 | 08:00-08:30 |
| Клиент 2, адрес 2 | 08:00-08:30 |
| Клиент 3, адрес 3 | 08:00-09:30 |
| Клиент 4, адрес 4 | 08:00-09:30 |
| Клиент 5, адрес 5 | 08:00-09:30 |
| Клиент 6, адрес 6 | 08:00-09:30 |
| Клиент 7, адрес 7 | 08:00-09:30 |
| Клиент 8, адрес 8 | 08:00-09:30 |
| Клиент 9, адрес 9 | 08:00-09:30 |
| Клиент 10, адрес 10 | 08:00-09:30 |

В таблице 2 показан результат, который выдает система при нормально настроенных временных окнах. На рисунке 1 изображен скриншот системы. Была сделана и сохранена в системе симуляция планирования на 17 июля. Настройки системы не были изменены за

исключением временных окон в системе. Если сравнить реальную доставку с той, что рассчитала система с правильными временными окнами, то видно, что система с небольшими отклонениями рекомендует тот же самый маршрут, что и в реальности проделал водитель. Следовательно, если все временные окна впоследствии настроить таким образом, то можно будет и объединять зоны доставки водителей, и в результате максимальный негатив, который будет получен от торговой точки, будет связан с приездом другого водителя, но претензий к времени доставки быть практически не должно, т. к. система строго следует тем временным интервалам, которые ей задают.

В Павлодаре на данный момент по категории ВС развозку производят 11 машин. При нормальном планировании они и будут развозить заказы по точкам на своей территории. Есть особенности павлодарской развозки, которых нет в других городах:

1. Самая главная особенность развоза молочной продукции в Павлодаре — это ежедневные заказы и ежедневная развозка товара по большому количеству точек. Приведем пример магазина «Ганза» на ул. Ткачева 12/7. Этот магазин заказывает продукцию каждый день, и его средний чек по анализу отчета по продажам около 20 тысяч тенге. Если руководствоваться самым простым методом, чтобы оптимизировать доставку в данный магазин, то необходимо уговорить ТТ, чтобы она заказывала товара в 2 раза больше (в среднем на 40 тысяч тенге), и в этом случае вывоз можно будет осуществлять не 6 раз в неделю, а лишь 3 раза. Но если проанализировать ситуацию со стороны самой торговой точки, то приходим к следующий выводам:

− Средний чек — 20000 тенге, это очень хороший заказ для ВС-категории;

− ВС-категории — это небольшие точки, где товар хранится обычно на полках или, в случае товара рассматриваемой компании, в холодильниках. У 95% ВС-категории нет даже небольшого склада, чтобы хранить запасной товар, который пользуется спросом. Все полки забиты до предела. Следовательно, даже если бы точка хотела заказать товара в 2 раза больше, чем обычно, этот товар просто негде будет хранить физически. Вследствие этого им удобно заказывать каждый день.

2. Если точка заказывает каждый день, значит, там происходит так, что и доставка, и торговый представитель компании приходит в точку каждый день (кроме воскресенья). Следовательно, необходимо доставить товар в точку как можно раньше, чтобы торговый представитель пришел после доставки. И это логично, т. к. маловероятно, что точка захочет делать заказ у торгового, когда товар, который был заказан еще вчера, не привезли.

Общие итоги по системе и тем преимуществам, которые она может предоставить отделу логистики в Павлодаре:

− для Павлодара система даст больше качественных преимуществ, нежели количественных, т. к. город небольшой и количество машин доставки, участвующих в планировании, также небольшое. Система может показать количественную эффективность и сэкономить 1–2 машины максимум, но при этом теряется взаимосвязь водитель-торговая точка, т. к. время отгрузки в точку знакомому водителю составляет существенно меньшее время, чем незнакомому, и незнакомый водитель еще должен потратить дополнительное время на поиск точки и правильную парковку. При переводе этих данных в сухую математику получается следующее: если в каждой точке доставки терять по 5 лишних минут и более, то суммарно это увеличит время доставки на 2–3 часа.

Общие выводы по работе филиала в Павлодаре с системой для распределения по категории ВС можно сделать следующие:

− необходимо как можно быстрее привести временные окна системы в порядок. Без участия компании сделать это невозможно, и до сих пор во многих точках неправильные временные окна доставки;

− только после приведения временных окон в порядок можно будет точечно корректировать дополнительные, более глубокие настройки, чтобы еще больше помочь в оптимизации маршрутов и повышении качества доставки.

| | Планируемый адрес | Временные окна операции | Планируемое время прибытия |
|---|---|---|---|
| | Калиакбаров Серик... (Foton 568AB14) 17.07.2019 001 | | |
| | FoodMaster Павлодар | ... | 06:00 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Ткачева, дом № 15/1 | 08:00-11:00 | 08:00 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Ткачева, дом № 12/7 | 08:00-11:00 | 08:09 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Ткачева, дом № 5 | 08:00-11:00 | 08:21 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Ткачева, дом № 7 | 08:00-11:00 | 08:26 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Майры, дом № 21/1 | 08:00-11:00 | 08:36 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Майры, дом № 27/1 | 08:00-11:00 | 08:43 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Майры, дом № 29 | 08:00-11:00 | 08:50 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Майры, дом № 41 | 08:00-11:00 | 08:56 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Майры, дом № 31 | 08:00-11:00 | 09:06 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Майры, дом № 49 | 09:00-12:00 | 09:15 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Бекхожина, дом № 3/3 | 09:00-12:00 | 09:22 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Бекхожина, дом № 15 | ... | 09:30 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Бекхожина, дом № 15/1 | 08:00-11:00 | 09:41 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Ткачева, дом № 9/1 | ... | 09:49 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Ткачева, дом № 3а | 09:00-12:00 | 10:04 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Затон, дом № 61 | 09:00-12:00 | 10:13 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.Чокина, дом № 25 | 09:00-12:00 | 10:22 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.1 Мая, дом № 284/1 | 09:00-12:00 | 10:31 |
| Д | Республика Казахстан, Павлодарская область, г.Павлодар, ул.1 Мая, дом № 284 | 09:00-12:00 | 10:37 |
| | FoodMaster Павлодар | | 10:49 |

*Рисунок 1 - Нормальные временные окна*

Рекомендация №1 по итогам проведенных наблюдений в Павлодаре: водители сейчас очень рано приезжают на склад за товаром (5.00–5.30), а точки открываются в 8.00, т. к. им после получения товара необходимо его самостоятельно распределить по ящикам, исходя из доставки, на что водители тратят от одного до двух часов, в зависимости от количества заявок. Возможно, чтобы сэкономить время водителей, стоит ввести в штат сотрудников, которые будут готовить товары для них с учетом распределения по торговым точкам. В качестве альтернативы можно доплачивать за эту работу водителям (они работают по одному, тогда как конкуренты, по наблюдениям, выходят на маршрут по двое, что существенно упрощает выполнение их обязанностей), т. к. предварительное распределение товара по точкам экономит время отгрузки в саму точку в 3–4 раза, а это существенная экономия времени.

Вот время работы водителя, если он приехал на погрузку к 5.30 утра:

− 5.30–6.00 — погрузка товара для доставки в около сорока точек;
− 6.00–6.20 — время поездки к первой точке доставки, которая откроется только в 8.00;
− 6.20–8.00 — распределение товара в кузове авто по точкам для удобства отгрузки;
− 8.00–13.00 — отгрузка в точки;
− 13.00–13.30 — возвращение на базу;
− 13.30–15.00 — сдача авто, отчетов, денег и т. д.

ИТОГО: водитель работает 9–10 часов.

**Выводы**

В последние годы транспортная логистика пережила много социальных, экономических и технологических изменений. Такие преобразования значительно влияют на всю логистическую цепочку от принятия заказа до распределения заказов клиентам. В связи

с растущим интересом к обеспечению эффективной доставки заказов клиентам, планирование стало важным аспектом компаний с логистическим отделом, а также для исследователей. В настоящей статье была рассмотрена проблема планирования в компании и пути ее решения, которые стали основой при реализации интеллектуальной системы принятия решений в логистике. Предлагаемая архитектура SOA на основе брокера использует модель прослеживаемости, чтобы предложить своим клиентам доступ к этим данным.

Система минимизирует человеческий фактор при планировании доставки за счет настроенной системы, т. е. планирование в перспективе может запускать любой человек, чьи знания компания посчитает достаточными для этого. Также она минимизирует человеческий фактор при планировании порядка доставки товара в точки (при правильно настроенных временных окнах, что для Павлодара очень важно).

Система сможет оперативно помочь в развозке в случае планируемого или внезапного отсутствия водителя (отгулы, больничные, отпуска и т. д.) или физической невозможности вывоза машиной (поломка, ремонт и т. д.).

С точки зрения качества, отдел логистики будет всегда получать оперативную информацию с полей о том, как доставляется товар, успевает ли водитель развести все вовремя и затем анализировать, выбирает ли водитель оптимальный маршрут или нет. На основе периодического анализа «планируемое / фактическое» можно дополнительно скорректировать настройки системы, т. к. экономия даже одной минуты на операции, которая повторяется 30–40 раз, — это экономия 30-40 минут времени планируемой доставки.

Система основана на потребностях в информации существующих организационных функций. В будущем система постарается задействовать инструменты, основанные на изменениях и информационных потребностях, которые помогут лицам, принимающим решения, адаптироваться в рабочей практике для удовлетворения будущих потребностей [8].

В будущем мы намерены расширить исследовательскую работу интеллектуальной системы принятия решений в логистике по личному кабинету водителей, по предоставленной обратной связи. Планирование будет дополнено математическим решением с учетом нескольких сценариев.

## СПИСОК ЛИТЕРАТУРЫ


1. Gorn A. P., Novikov D. T., Subbotin A. S. Usloviya innovaczionnogo razvitiya e`konomiki Rossii. – Federal`noe gosudarstvennoe avtonomnoe obrazovatel`noe uchrezhdenie vy`sshego obrazovaniya «Tyumenskij gosudarstvenny`j universitet», 2012. – S. 307.
2. Kolesnikov A. V., Kirikov I. A. Metodologiya i tekhnologiya resheniya slozhny`kh zadach metodami funkczional`ny`kh gibridny`kh intellektual`ny`kh sistem. – 2007. – C. 2–4.
3. Stankovskaya A. I. Organizacziya logisticheskogo upravleniya zapasami na predpriyatii. – 2019. - S. 7.Leedy P. D., Ormrod J. E. Practical research. – Saddle River, NJ : Pearson Custom, 2005. – V. 108. – № 13. – P. 347-355.
4. Creswell J. W., Creswell J. D. Research design: Qualitative, quantitative, and mixed methods approaches. – Sage publications, 2017. – P. 13.
5. Williams C. Research Methods //Journal of Business & Economic Research (JBER). – 2007. – V. 5. – № 3. – P. 11-17.
6. Ry`zhkova N. G., Aksenov K. A., Nevolina A. L. Analiz informaczionny`kh sistem podderzhki prinyatiya reshenij v sfere logistiki //Sovremenny`e problemy` nauki i obrazovaniya. – 2014. – #. 6. – S. 4-4.
7. Pashaev M. Ya. Sistema prinyatiya reshenij v ramkakh konczepczii ispol`zovaniya GLONASS dlya otslezhivaniya ob`ektov v zone kontrolya //Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Seriya: Upravlenie, vy`chislitel`naya tekhnika i informatika. – 2017. – #. 2. – S. 89-92.Gold S., Kunz N., Reiner G. Sustainable global agrifood supply chains: exploring the barriers //Journal of industrial ecology. – 2017. – V. 21. – №. 2. – P. 249-260.

8. Kamariotou M., Kitsios F. Strategic information systems planning: SMEs performance outcomes //Proceedings of the 5th International Symposium and 27th National Conference on Operation Research. – 2016. – P. 153-157.
9. Karthik B. et al. Decision support system framework for performance-based evaluation and ranking system of carry and forward agents //Strategic Outsourcing: An International Journal. – 2015. – P. 11-15.

**Адырбек Ж.А., Сатыбалдиева Р.Ж.**

**Логистикадағы жоспарлау процестерін талдау және логистикадағы интеллектуалды жүйені қолдану арқылы мәселелерді шешу**

**Аңдатпа.** Интеллектуалды жүйемен компания қалалық тарату маршруттарын жоспарлауға және бензин мен жалақыға ақшаны үнемдейді. Жүйе FMCG секторына, интернет-дүкендерге, сатушыларға, дистрибьюторларға және көлік логистикасы бөлімі бар барлық басқа салаларға арналған. Шешімді 20 машинасы бар шағын компаниялар да, әлдеқайда көп машиналары бар ірі компаниялар да қолдана алады. FMCG жүйесінің көмегімен компания тауарларды жеткізу құнын екі есеге төмендетуге мүмкіндігі бар. Көптеген компаниялар логистикалық қызметкерлердің жұмыс уақытын оңтайландыра алмағандықтан жылына миллиондаған қаражат жұмсайды. Ірі компанияларда жұмыстың ұйымдастырылуы, процесте проблемалар туындаған жағдайда, тығырықтан шығу жолдарын жақсартуға болатындығын түсіну мақсатында зерттеу жүргізілді. Деректер жиналды, процеске сәйкес талданды, содан кейін маршрут салынды, осы маршруттар бойынша курьерлер жіберілді және қайтадан тағы талдау жүргізілді.

**Түйінді сөздер:** интеллектуалды көліктік логистиканы автоматтандыру жүйесі, көлік логистикасын автоматтандыру, көлік логистикасындағы бизнес-процестерді автоматтандыру, логистикадағы интеллектуалды жүйенің бизнес-процестің моделі.

**Adyrbek Zh.A., Satybaldiyeva R.Zh.**

**Analysis of the planning and problem-solving processes in logistics using an intelligent system**

**Abstract.** An intelligent system helps a logistics company to save time on planning urban distribution routes and money on gasoline and salaries. The system is intended for the FMCG sector, online stores, retailers, distributors, and all other sectors which have a transport logistics department. The solution can be used by both small companies, which have up to 20 cars, and large companies, which have much more cars. With the help of the FMCG system, the company can halve the cost of shipping goods. Many companies annually spend millions on it because they cannot optimize the working hours of their logistics staff. The articles presents the results of the research into the processes' arrangement in large companies, the problems emerging in the process, and ways to improve the situation. In the course of our research we collected and analyzed data on the process, then built the most optimal route, launched couriers along these routes and performed the analysis again.

**Keywords:** intelligent system, automation of transport logistics, transport logistics automation, automation of business processes in transport logistics, business process model of an intelligent system in logistics.

**Авторлар туралы мәлімет:**

**Адырбек Жансая Адырбекқызы**, Халықаралық ақпараттық технологиялар университеті, «Ақпараттық жүйелер» кафедрасының магистранты.

**Сатыбалдиева Рысхан Жакановна**, техника ғылымдарының докторы, Халықаралық ақпараттық технологиялар университеті, «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры.

**Сведения об авторах:**

**Адырбек Жансая Адырбекқызы**, магистрант кафедры «Информационные системы», Международный университет информационных технологий.

**Сатыбалдиева Рысхан Жакановна**, кандидат технических наук, ассоциированный профессор кафедры «Информационные системы», Международный университет информационных технологий.

**About the authors:**

**Zhansaya A. Adyrbek**, master student, Department of Information Systems, International Information Technology University.

**Ryskhan Zh. Satybaldiyeva**, PhD in Engineering Science, Associate Professor, Department of Information Systems, International Information Technology University.

УДК 001.8, 008, 005.94

**Nurgaliyev M.K.[1], Alimzhanova L.M.[2]**

[1,2] International Information Technology University, Almaty, Kazakhstan

# GAMIFICATION IN EDUCATION

**Abstract.** The rapid advancement of technologies and the growing amount of information exchange provide new approaches to improving educational processes through various methods. Gamification becomes one of those effective practices and stimulations to increase motivation and engagement of learners in the educational environment. Given the importance of education in contemporary academics and culture, this article examines how gamification is being used in the learning process and provides a more precise picture of the results achieved so far and how they were achieved.

**Keywords:** Gamification, Game-based Learning, Education, E-learning, game mechanics, game dynamics, game design.

## Introduction

It's a proven fact that a country's development crucially depends on human resources. As it is known, education is responsible for shaping a person; thus, the educational system plays an important role in building a foundation for the nation's prosperity. With the higher literacy rate, there is a reduced percentage of unemployment. To build a strong educational system, there should be effective pedagogical methods, involving the rapidly advancing social networks and technologies which have become an everyday routine of human life. With the emergence of new blogs, services, social networking platforms and websites they are being frequently used for knowledge purposes nowadays. Similarly, back to the early 2000s, computer adaptive games were used for learning second languages at some schools [1].

But now with the development of high technologies such as the Internet, embedded systems (IOT), mobile gadgets, education is evolving even faster. With the massive use of high technology, people are beginning to search more effective methods and tools for more effective teaching of both children and adults. One of these methods is the gamification of educational processes. Gamification is described by scientists in a variety of ways. But they all share the same opinion that gamification is a non-game process in which elements of a game nature are included. These game elements include the scoring system, leaderboards, in-game gifts, interesting stories, and others. They are used to increase the motivation, engagement and quality of student learning, not only in the educational environment, but also in other areas of life. The use of gamification in the educational sector is gaining popularity, gradually proving its effectiveness. As a result, many educators, educational institutions and even large companies are incorporating elements of gamification into their learning processes. The structure and design of games allows of implementing elements of gamification within the learning method, and of an intuitive interpretation of the mechanics of game processes.

## Definitions

According to Kapp [2] gamification is "the use of game theory, graphics and game-based mechanics to encourage learning, inspire action, involve users and solve problems".

Gamification is the use of game thinking, approaches, and elements in a non-game sense. In both formal and casual settings, using game mechanics increases inspiration and understanding, in other words gamification is the application of game elements and game thinking in non-game practices. In gamification, games have some distinct significant characteristics such as:

- Users are employees or customers (for businesses), students (for educational institutions);

- users complete and advance challenges/tasks against specified objectives;
- points are earned as a result of performing tasks;
- users pass through levels based on the points earned;
- the granted awards function as incentives for completing actions;
- users are ranked according to their accomplishments.

**Game dynamics and game mechanics**

Games that are entertaining include enjoyable experiences, and it appears that, far from waning, interest in recreational games is still increasing. Fitocracy, Runkeeper, Nike+, Zombies, Run! and other computer-assisted gamified services help to structure, endorse, and motivate fitness practices [3].

It's also been proposed that commercial game players improve their problem-solving and reading abilities, and that successful commercial games embody good learning values by allowing gamers to participate effectively and reflectively while playing [4].


*Figure 1 - The game economy*

The agents, items, components, and their relationships in the game are referred to as the game mechanics. They describe the game as a rule-based structure that specifies what is present, how it functions, and how players can communicate with the game environment. Game dynamics is the emergent behavior that occurs as the mechanics are used in a game, and aesthetics are the players' emotional responses to the game play [5].

Points, ranks, medals, trophies, virtual items, leader boards, and virtual gifts are all well-known game mechanics. Rewards, rank, competition, self-expression, and others are game dynamics components, according to Schonfeld. He depicts 47 different game dynamics elements [6].

**Gamification in education**

An effective education environment can encourage contact with students and teachers, reciprocity, fast feedback, collaboration among students, constructive learning strategies, time on mission, appreciation for diversity, communication of high expectations and students' different learning styles [7].

The key goal of education is to increase student performance, productivity, commitment, happiness, and inspiration. The use of game mechanics and gamification will help accomplish these goals.

Education management is a crucial component of the model which helps students to be inspired, happy, effective, and efficient. The model shown in Figure 2 includes the following key elements: management, game dynamics, critical factors, development phases, user experience elements, game mechanics, gamification elements, and their impact on students.

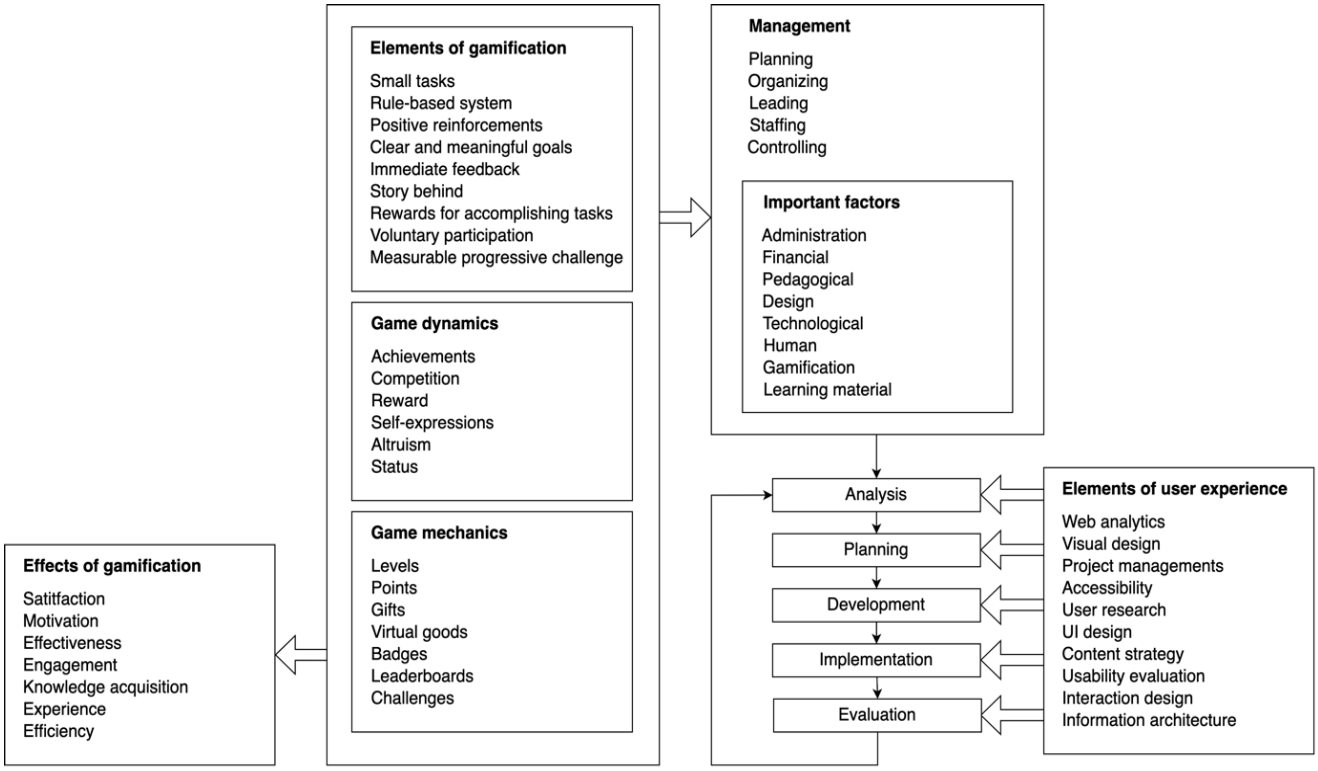



*Figure 2 - The model of gamification in education*

**Impact of instructional content on learning outcomes**

The lack of high-quality learning resources along with gamification would not improve learning outcomes. Gamifying the education would not aid in success, if the material of education does not assist students in learning [8]. This statement serves as a useful note for academics. High-quality instructional material is needed regardless of technological or pedagogical advances. Student learning experiences may be significantly influenced by attitudes and behaviors. While the constructs (e.g., innate motivation) can differ across contexts, evidence of significant relationships between student attitudes and behaviors and student learning outcomes has been found in the educational literature [9]. Students who put in more cognitive effort, spend more time on task, and have a constructive outlook toward a subject can theoretically see better results [10].

**Conclusion**

The results of this study confirm the hypothesis that gamification is gaining momentum in education, especially among the younger generation as it helps to motivate and involve students in learning.

This study also shows that gamification is not a panacea for everything, and it should be used only when necessary.

REFERENCES

1. Dervis Kayımbaşıoğlu, Procedia Computer Science 102, 2016, 668-676 p.
2. K. M. Kapp. The gamification of learning and instruction: game-based methods and strategies for training and education, Pfeiffer, 2012, 55-63 p.
3. Hamari, J., & Koivisto, J. Social motivations to use gamification: An empirical study of gamifying exercise. In Proceedings of the 21st European conference on information systems, 2013, 54-59 p.
4. Gee, J. P. What video games have to teach us about learning and literacy, Palgrave Macmillan, 2003, 34-47p.
5. Grünberg T. K. What's the difference between game mechanics and game dynamics? [Electronic resource] URL: https://www.quora.com/What-is-the-difference-between-game-mechanics-and-game-dynamics, (retrieving date: 21.04.2021).
6. Schonfeld E. SCVNGR's Secret Game Mechanics Playdeck. [Electronic resource] URL: https://techcrunch.com/2010/08/25/scvngr-game-mechanics/, (retrieving date: 22.04.2021).
7. Shea P. J., Pickett A.M., Pelz W.E. A follow-up investigation of teaching presence in the SUNY learning network. Journal of Asynchronous Learning Networks, 7(2), 2003, 61–80.
8. Bedwell, W. L., Pavlas, D., Heyne, K., Lazzara, E. H., & Salas, E. Toward a taxonomy linking game attributes to learning: An empirical study. Simulation & Gaming, 43(6), 2012, 729–760p.
9. Hattie, J. Visible learning: A synthesis of over 800 meta-analyses relating to achievement, Routledge, 2008, 49-55 p.
10. Richardson, M., Abraham, C., & Bond, R. Psychological correlates of university students' academic performance: A systematic review and meta-analysis. Psychological Bulletin, 138(2), 2012, 353–387p.

**Нұрғалиев М.Қ., Алимжанова Л.М.**

**Білім беру саласындағы геймификация**

**Аңдатпа.** Технологияның қарқынды дамуы және ақпарат алмасудың өсіп келе жатқан көлемі әртүрлі әдістер арқылы білім беру процестерін жақсартуға жаңа мүмкіндіктер ашады. Геймификация оқушылардың білім беру ортасына деген ынтасы мен белсенділігін арттыратын тиімді тәжірибелер мен ынталандырулардың біріне айналуда. Қазіргі ғалымдар мен мәдени топтар үшін білімнің маңыздылығын ескере отырып, бұл мақала геймификацияның оқу процесінде қолданылатыны жайында зерттейді және басқа еңбектерде жазылған нәтижелер мен оларға қол жеткізілгені туралы нақты түсінік береді.

**Түйінді сөздер:** геймификация, ойын арқылы білім беру, білім беру саласы, электронды оқыту, ойын механикасы, ойын динамикасы, ойын дизайны.

**Нургалиев М.К., Алимжанова Л.М.**

**Геймификация в образовании**

**Аннотация.** Быстрое развитие технологий и растущий объем обмена информацией открывает новые возможности для улучшения образовательных процессов с помощью различных методов. Геймификация становится одной из эффективных практик и стимулом, которые повышают мотивацию учащихся и их вовлеченность в образовательную среду. В этой работе исследуется, как геймификация используется в процессе обучения, дается более точное представление о результатах, зафиксированных в статьях, и о том, как они были достигнуты.

**Ключевые слова:** геймификация, игровое обучение, образование, электронное обучение, игровая механика, игровая динамика, игровой дизайн.

**Авторлар туралы мәлімет:**

**Нұрғалиев Медет Құрманғазыұлы**, «Ақпараттық жүйелер» кафедрасының 2-ші курс магистранты, Халықаралық ақпараттық технологиялар университеті.

**Алимжанова Лаура Муратбековна**, т.ғ.к, «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Нургалиев Медет Курмангазыулы**, магистрант 2 курса кафедры «Информационные системы», Международный университет информационных технологий.

**Алимжанова Лаура Муратбековна**, к.т.н., ассоциированный профессор кафедры «Информационные системы», Международный университет информационных технологий.

**About the authors:**

**Medet K. Nurgaliyev**, master student, Department of Information Systems, International Information Technology University.

**Laura M. Alimzhanova,** Candidate of Technical Sciences, Associate Professor, Department of Information Systems, International Information Technology University.

УДК 004.5

**Базарбеков И.М.[1], Шарипов Б.Ж.[2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан

# РАЗРАБОТКА БИЗНЕС-ПРОЦЕССА ДЛЯ ПОЛУЧЕНИЯ ОНЛАЙН УСЛУГ В ОРГАНИЗАЦИИ ОБРАЗОВАНИЯ

**Аннотация.** В данной статье рассмотрены и разработаны процессы получения онлайн услуг в организациях образования на примере университета АО «МУИТ». Под онлайн услугами подразумеваются такие услуги, как: получение справки с места обучения, получение транскрипта, подача заявлений онлайн, ликвидация задолженности. Были разработаны карта процессов и модели нотации BPMN «AS IS» и «TO BE».

**Ключевые слова:** “Smart campus”, Business Process Model and Notation, “AS IS”, “TO BE”, онлайн услуги, система, электронная цифровая подпись.

Современные университеты имеют разнообразную и большую инфраструктуру. Университет занимается не только образовательным процессом, но и административно-организационными составляющими. Если для образовательного процесса требуется всего одна система дистанционного обучения (moodle), то функционал для административно-организационных процессов может быть весьма разнообразен. В условиях карантина без подобных систем не обойтись. Ключевым звеном университетской инфраструктуры является кампус университета как коммуникативная среда взаимодействия студентов, докторантов, преподавателей и научных работников, что является неотъемлемой составляющей учебного процесса. В соответствии с доминирующими тенденциями, определяющими развитие современной системы образования и связанными с внедрением новых информационных технологий и формированием единого научно-образовательного пространства, коммуникативная среда кампуса должна базироваться на применении современных ИТ-решений. В такой постановке университетский электронный кампус («Smart campus») становится важным инфраструктурным элементом с полным циклом автоматизации важнейших задач деятельности университета, предоставлением персонализированного информационного пространства и соответствующих информационных услуг.

В настоящее время такие услуги, как получение справки с места обучения в АО «МУИТ», получение транскрипта, ликвидация задолженности, подача заявлений в деканат, не автоматизированы в полной мере. Например, процесс получения справки с места обучения происходит следующим образом: обучающийся отправляет на почту заявку с указанием ФИО, группы и курса обучения. После этого, в течение трех рабочих дней, студент забирает документ из университета либо получает отсканированную копию на почту. Однако сотрудники университета сканируют или вносят изменения в справки вручную. Также больших временных затрат требует подписание справки руководством университета. Остальные процессы, такие, как подача заявлений, ликвидация академических задолженностей, происходят через почту и обрабатываются вручную, на регистрацию каждого письма и отправки ответа уходит большое количество времени. Система «Smart campus» позволит автоматизировать данные процессы и будет генерировать справки, регистрировать входящие номера для заявлений моментально, сразу после подачи заявки студентами, что значительно сократит нагрузку на административный персонал и улучшит условия для обучающихся.

**Моделирование процессов**

*А. Карта процессов*

Карта процессов «Получение онлайн услуг» - инструмент планирования и управления процессом, который используется для повышения эффективности выполняемых действий в течении всего процесса. Карта процессов отражает следующие виды процессов:

- Основные процессы.
- Процессы управления.
- Вспомогательные процессы [1].

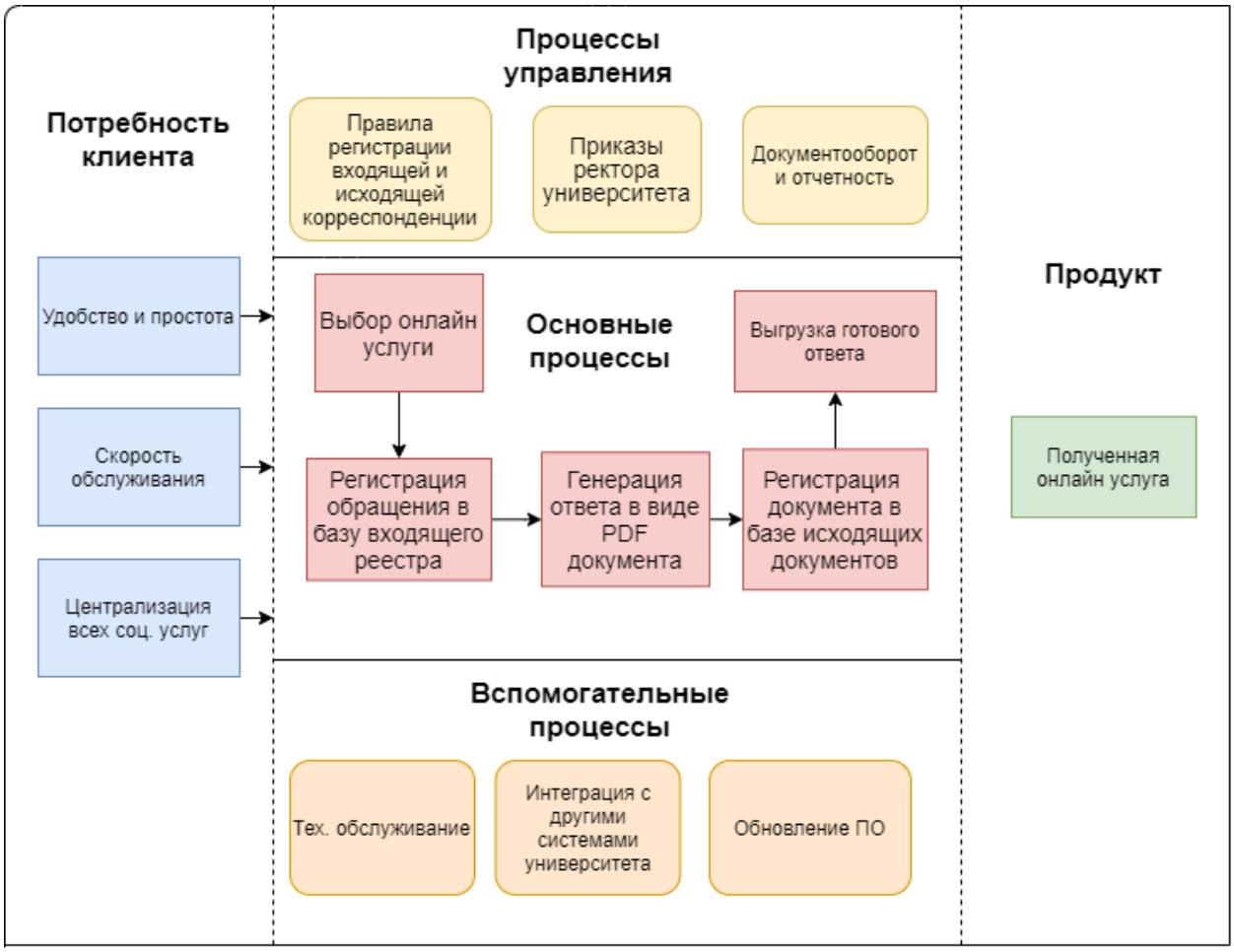



*Рисунок 1 - Карта процесса «Получение онлайн услуг в системе "Smart-campus"пан»*

На рисунке 1 приведена карта процесса «Получение справки с места обучения». На карте процесса показаны следующие процессы:

Основные процессы:

- Выбор онлайн услуги;
- Регистрация обращения в базе входящего реестра;
- Генерация ответа в виде PDF документа;
- Регистрация номера выданного документа в базе исходящих документов;
- Выгрузка готового ответа.

Процессы управления:

- Правила регистрации входящей и исходящей корреспонденции;
- Документооборот и отчетность;
- Приказы ректора университета.

Вспомогательные процессы:

- Тех. обслуживание;
- Обновление ПО;
- Интеграция с другими системами университета.

*B. BPMN модель бизнес-процессов*

BPMN модель включает в себя набор концепций, необходимых для моделирования процессов. BPMN модель является связующим звеном между техническими разработчиками и бизнес-пользователями процесса. Модель также отражает все существующие интеграции и связи различных систем [2]. Ниже, на рисунке 2, приведена BPMN модель процесса получения онлайн услуг в системе «Smart campus».

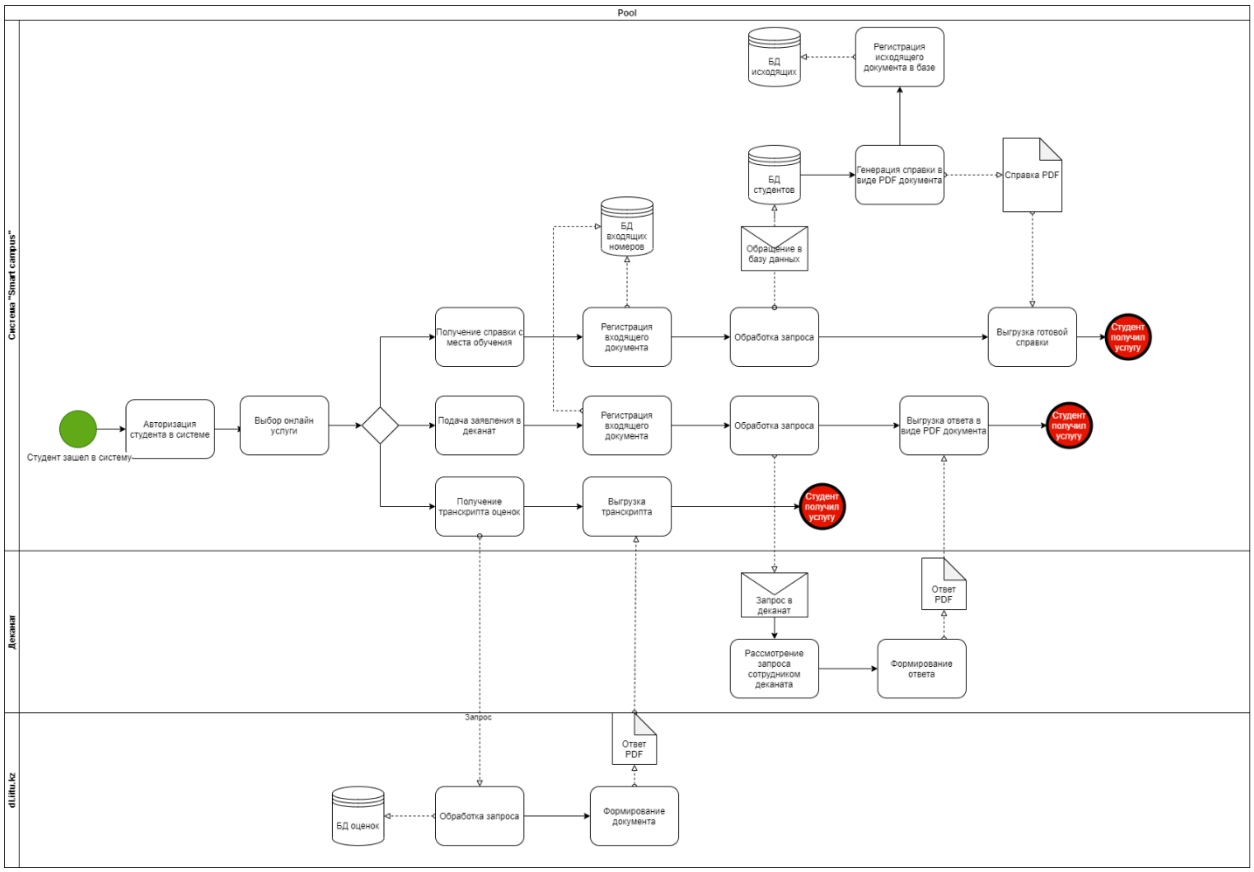



*Рисунок 2 - BPMN модель процесса «AS IS»*

**Регламент бизнес-процесса**

Цель процесса — автоматизировать и упростить порядок получения справок, транскриптов, отчетов по оплате, подачи заявлений и других услуг посредством их получения в режиме онлайн, без ожидания своей очереди на прием к ответственным за это людям. Помимо облегчения процесса для обучающихся автоматизация получения данных услуг экономит ресурсы персонала университета, значительно сокращая их задачи. Владельцем процесса является университет в лице директора департамента технического сопровождения и IT-поддержки.

Владелец процесса несет ответственность за:

- бесперебойную работу системы в любое время суток;
- быстрое реагирование на запросы технических неполадок и их решение;
- своевременное обновление данных и инструментов системы;
- интеграцию системы с другими департаментами и внутренними системами;
- выполнение процесса и его результат.

Обычно определение подхода к управлению процессами включает следующие этапы:

- Документирование процесса для понимания того, как работа проходит через процесс;
- Присвоение права собственности на процесс с целью установления управленческой подотчетности;
- Управление процессом для оптимизации некоторых показателей эффективности процесса;
- Улучшение процесса для повышения качества продукции или показателей эффективности процесса;
- Управляя процессами, можно лучше интегрировать перспективы и приоритеты с ресурсами;
- Многие новые управленческие инициативы требуют управления процессами, и их невозможно реализовать;
- Управление процессами открывает двери для творческих и новаторских подходов к улучшению организационной деятельности;
- Управление процессами позволяет эффективно внедрять современные системы и стандартное программное обеспечение [3].

Можно сделать вывод, что для поддержки инноваций в данном процессе необходим систематический подход к проектированию и анализу. Ниже, в таблицах 1 и 2, показаны описание процесса получения онлайн услуг и характеристика данного процесса.

Таблица 1 –Описание процесса в системе “Smart-campus”

| **Название процесса** | **Тип процесса** | **Цель / назначение процесса** | **Владелец процесса** |
|---|---|---|---|
| Получение онлайн услуг в системе “Smart campus”. | Основной | Упростить порядок получения справок, транскриптов, отчетов по оплате, подачи заявлений и других услуг посредством их получения онлайн, не ожидая очереди на прием к ответственным за это людям. | ВУЗ, администрация ВУЗа, департамент технического сопровождения и IT-поддержки. |

Таблица 2- Характеристика процесса в системе “Smart-campus”

| **Основное событие начала** | **Основной вход** | **Основной поставщик** | **Основное событие окончания** | **Основной продукт** | **Основной клиент** |
|---|---|---|---|---|---|
| Регистрация в системе “Smart campus” / выбор онлайн услуги | 1) Запрос на получение онлайн услуги (получение справки, транскрипта, информация об оплате);<br>2) Данные клиента(студента) | Система “Smart-campus” | Клиент(студент) получил онлайн-услугу (готовый документ) | Справка об обучении, транскрипт, отчет по оплате | Обучающийся в организации образования (ВУЗа) |

**Модель бизнес-процессов «ТО BE».** Модель бизнес-процессов «ТО BE», то есть «как должно быть» отражает оптимизированные и доработанные аспекты модели «как есть»

[4]. Перед построением данной модели «как должно быть» необходимо исследовать плюсы и минусы модели «как есть», проанализировать возможные варианты улучшения эффективности процессов и выбрать подходящие методы решения данных проблем. Ниже, в таблице 3, приведены основные проблемы и их решения.

Таблица 3 – Проблемы модели «как есть»

| Проблемы | Решения |
|---|---|
| Сбой системы | 1. Принимаются меры по устранению департаментом технического сопровождения и IT поддержки.<br>2. В случае затяжных технических работ необходимо прописать в инструкции, что заявка подается альтернативными способами (e-mail, whats app и т.д.), затем заявка обрабатывается вручную сотрудником деканата. |
| Проверка подлинности документа | Для того чтобы справка имела юридическую силу, необходимо внедрить ЭЦП, а для проверки подлинности в углу документа генерировать QR код, который откроет с устройства электронный документ по адресу официального сайта университета [5]. |
| Аутентификация пользователя | Добавление номера телефона и смс-подтверждение, а также внедрение ЭЦП. |

После исследования всех элементов модели «как есть» BPMN модель бизнес-процессов «Получение онлайн услуг в системе “Smart campus”» была оптимизирована. Модель после оптимизации «TO BE» («как должно быть») показана ниже, на рисунке 3.

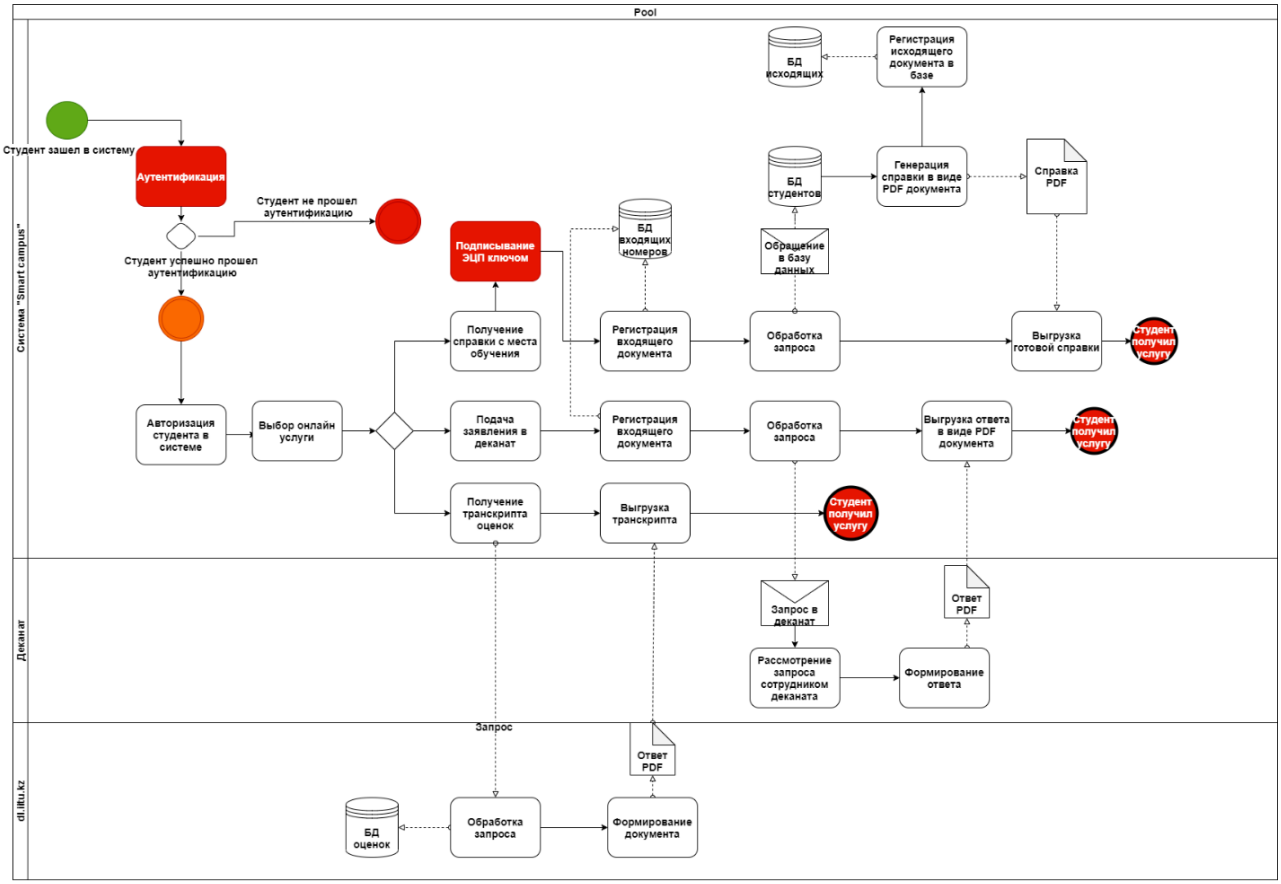



*Рисунок 3 - BPMN модель процесса «TO BE»*

В заключении отметим, что в данной статье была разработана модель бизнес-процессов получения онлайн услуг в система «Smart campus», также разработанная модель была оптимизирована после исследования и анализа. Были описаны основные бизнес-процессы предоставления онлайн услуг в университете, выявлены проблемы существующих моделей и предложены пути их решения.

СПИСОК ЛИТЕРАТУРЫ

1. Веб-приложение учета и регистрации абитуриентов АО «МУИТ»: система «CAMPUS IITU», [Электронный ресурс] URL: https://campus.iitu.kz. (дата обращения: 10.04.2021)
2. BPMN 2.0 ИЗ ЧЕГО СОСТОИТ МОДЕЛЬ БИЗНЕС ПРОЦЕССА, [Электронный ресурс] URL: https://rzbpm.ru/knowledge/bpmn-2-0-iz-chego-sostoit-model-biznes-processa.html (дата обращения: 11.04.2021)
3. Регламент бизнес-процесса. Глоссарий, [Электронный ресурс] URL: https://www.businessstudio.ru/articles/article/glossariy_reglament_protsessa (дата обращения: 15.04.2021)
4. Портал СМК ВКТУ, [Электронный ресурс] URL: https://www.ektu.kz/abouttheuniversity/qms/processmap.aspx?lang=ru (дата обращения: 15.04.2021)
5. Портал электронного правительства - EGOV,[Электронный ресурс] URL: https://www.egov.kz (дата обращения: 15.04.2021)

REFERENCES

1. *Veb-prilozhenie ucheta i registracii abiturientov AO "MUIT": sistema "CAMPUS IITU"*, [The web application of accounting and registration of enrollees of JSC "IITU"], [Electronic resource] URL: https://campus.iitu.kz (accessed: 10.04.2021)
2. *BPMN 2.0 IZ CHEGO SOSTOIT MODEL BIZNES PROTSESSA*, [BPMN 2.0 WHAT IS A BUSINESS PROCESS MODEL] [Electronic resource] URL: https://rzbpm.ru/knowledge/bpmn-2-0-iz-chego-sostoit-model-biznes-processa.html. (accessed: 11.04.2021)
3. *Reglament biznes-protsessa. Glossariy*, [Business process regulations. Glossary], [Electronic resource] URL: https://www.businessstudio.ru/articles/article/glossariy_reglament_protsessa (accessed: 15.04.2021)
4. Portal SMK VKTU, [SMK VKTU portal], [Electronic resource] URLhttps://www.ektu.kz/abouttheuniversity/qms/processmap.aspx?lang=ru (accessed: 15.04.2021)
5. Portal elektronnogo pravitelstva - EGOV, [E-government portal- EGOV], [Electronic resource] URL: https://www.egov.kz (accessed: 15.04.2021)

**Базарбеков И.М., Шарипов Б.Ж.**

**Білім беру ұйымында онлайн қызмет көрсету үшін бизнес-процессін дамыту**

**Аңдатпа.** Бұл мақала «ХАТУ» АҚ Университетінің мысалында білім беру ұйымдарында онлайн-қызметтерді пайдалану процестерін дамыту туралы. Онлайн-қызметтер дегеніміз: оқу орнынан сертификат алу, транскрипт алу, өтінімдерді онлайн режимінде беру, қарызды жою. Технологиялық карта және BPMN «AS IS» және «TO BE белгілеу модельдері» әзірленді.

**Түйінді сөздер**: «Smart-кампус», Business Process Model and Notation, "AS IS", "TO BE", онлайн қызмет көрсету, жүйе, электрондық цифрлық қолтаңба.

**Bazarbekov I.M., Sharipov B.Zh.**

**Development of a business process for obtaining online services in the organization of education**

**Abstract.** This article presents a case study of IITU JSC illustrating expansion in the use of online services at educational organizations. Online services mean such services as: obtaining a certificate from the place of study, receiving a transcript, submitting applications online, eliminating debt. To streamline the process of rendering such services online we developed a process map and some BPMN notation models such as "AS IS" and "TO BE".

**Keywords:** Smart-campus, Business Process Model and Notation, “AS IS”, “TO BE”, online services, system, electronic digital signature.

**Авторлар туралы мәлімет:**

**Базарбеков Икрам Медеуұлы** «Ақпараттық жүйелер» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті.

**Шарипов Бахыт Жапарович,** т.ғ.к., п.ғ.д, «Ақпараттық жүйелер» кафедрасының профессоры, білім беру инновациясы және SMART оқыту орталығының директоры, ХИНА академигі, РЖҒА шетелдік академигі, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Базарбеков Икрам Медеуұлы,** магистрант кафедры «Информационные системы», Международный университет информационных технологий.

**Шарипов Бахыт Жапарович,** к.т.н., д.п.н., профессор кафедры «Информационные системы», академик МАИН, иностранный академик РАЕН, директор центра образовательных инноваций и SMART обучения Международного университета информационных технологий.

**About the authors:**

**Ikram M. Bazarbekov,** master student, Department of Information Systems, International Information Technology University.

**Bakhyt Zh. Sharipov,** Dr of P.Sci, Cand. of Techn.Sci. , academician of IAIN, foreign academician of RANS, Director of the Center for Educational Innovation and SMART Training, Professor, Department of Information Systems, International Information Technology University.

УДК 53.06

**Жунусов Д.О.[1], Алиаскаров С.Ж.[2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан

# МЕТОД КЛАССИФИКАЦИИ ТЕКСТОВ НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

**Аннотация**. В этой работе были описаны способы расчета и анализа метода классификации текстов, а также были выявлены основные параметры применения алгоритмов машинного обучения для обработки естественных языков. Для этого были проанализированы и показаны некоторые реализации алгоритмов машинного обучения. Для разработки системы по анализу и категоризации текстов были применены алгоритмы по обработке текстовых данных, а также результаты исследования системы обработки естественного языка.

**Ключевые слова:** обработка, естественные языки, классификация, анализ, категоризация, распознавание, понимание.

## Введение

В цифровом мире важность обработки информации возрастает с каждым днем. Быстрое развитие технологий и разработка систем по мониторингу социальных сетей определили одно из ключевых направлений деятельности индустрии ИТ — создание систем по обработке естественного языка, которые могут обрабатывать любые виды текстовой информации: как новостные публикации и статьи, так и их комментарии. На сегодняшний день это проявляется в виде таргетированных предложений, составления портрета пользователя и прогнозирования действий на основе понимания текстов. Одной из самых востребованных технологий являются алгоритмы, основанные на NLP (natural language processing), поскольку данная сфера еще мало изучена. Обработка естественного языка — область, которая возникла в результате синхронизации таких наук, как лингвистика и математика [1]. Но одной классификации для точного понимания текста недостаточно. Необходимо категоризировать текст для дальнейшей обработки. Классификация текстов — это процесс синтаксического анализа текстов в различных сферах для дальнейшей работы с данными [2]. Далее следует визуализация данных и их предобработка для принятия решений в информационных системах. Однако для классификации текстов необходимы точные алгоритмы машинного обучения и большие корпусы размеченных данных, которые создают риски потери семантического значения категорий. Поэтому для обработки текстов, в частности для классификации текстов по категориям, требуется более эффективный метод анализа. Использование комплексных, более сложных методов анализа текстов может уменьшить вероятность ошибок, но повлечет увеличение требований к ресурсам. Предполагается, что есть необходимость в повышении точности и использовании более продвинутых алгоритмов для классификации.

## Методология

В данной статье будут рассматриваться методы классификации текстов, поэтому объектом исследования является анализ текста. Не существует стандартизированных шаблонов для анализа текстов. Анализ текста является примером области междисциплинарного исследования, соответственно, понятие анализа трансформируется в зависимости от области и от того, каков общий контекст статьи, связанной с данным понятием.

Классификация — довольно распространенная задача в машинном обучении. Она исследуется в таких сферах, как: распознавание и обработка изображений, сегментация объектов, нахождение дистанций и параметров, медицинские задачи и проблемы науки. Специалисты машинного обучения, работающие над сложными алгоритмами нейронных сетей, стремятся обеспечить их необходимое функционирование в определенных, заранее неизвестных условиях, то есть добиваются от разрабатываемых систем и алгоритмов необходимых форм поведения. Лингвисты изучают информацию, определяющую составное функционирование текста, а также единицу каждого слова в нужном контексте. Ключевыми функциями когнитивной лингвистики являются особенности усвоения и обработки информации [3]. Условно говоря, обработка текстовой информации — это нахождение последовательностей в текстах, поступивших из различных информационных источников. Это определение обработки информации в данной статье взято за основу, то есть текст рассматривается как единица вычисления, а классификация текстов подразумевает анализ того, как они себя показывают в разных условиях. Вслед за определением понятий нужно также выявить измеримые и высчитываемые параметры. В данном случае это контекст, в котором единицы участвуют для дальнейшей предобработки.

В рамках данного исследования метод классификации выступает как основа для дальнейших взаимодействий в системе, дополняя систему в качестве полноценного цикла понимания текста. Поэтому первичным действием, определяющим категорию текста, является сегментация. Каждая единица текста имеет определенный вес и будет рассматриваться в рамках всего контекста, в котором будут выстраиваться связи.

Материал для изучения был получен методом выгрузки корпуса больших данных по заметкам в Википедии и данных, находящихся в открытом доступе. В общей сложности было получено 600,000 оригинальных текстов. Для успешной реализации, поставленной задачи корпус данных был расширен методом исследования социальных сетей для выгрузки наибольшего числа оригинальных словосочетаний и предложений. Анализ сплошного содержания заметок был проведен методом выявления главных тем, методом прибавления первичного корпуса тегов исследования. Теги исследования задавались для поиска по социальным сетям и по новостным порталам. Роль человека в сборе была минимальной. В корпус больших данных методом оценки уникальности добавлялись предложения и создавали кластеры оригинальных услуг. Основная масса данных, собранных по тегам, была на русском и казахском языках. Все наборы данных не форматировались и не обрабатывались в процессе сбора, и в корпус могли попадать стоп-слова и малозначимые тексты. В общей сложности были задействованы программы для сбора данных из двенадцати социальных сетей и восьмисот новостных порталов, что позволило повысить качественный и количественный состав корпуса для дальнейшей обработки.

Для сбора данных из открытых источников использовалась техника «поиск-ключ», которая работает на основе языка python. Эта техника включает в себя инструменты для выгрузки html варианта страницы и поиска по тегам. Теги, соответствующие теме, были получены через поисковые системы методом поиска внутри источников.

Для сбора данных из новостных порталов был применен поиск по темам и дальше по тегам. Портал выгружался отдельным скриптом через определенные периоды времени и сохранялся в корпус для последующей обработки. Поиск слов проводился по новостным публикациям и по комментариям под ними путем установления фильтрации по времени (от новых к старым).

В общей сложности было собрано 21,231 текстов по социальным сетям и новостным порталам. Например, проводился поиск по тегам на политические и общественные темы, статистика обрисовывала настроение населения в определенный отрезок времени.

Классификации по типу были разбиты на несколько частей по функциям и базисным тегам. При делении на категории учитывались смежные элементы. Впоследствии выявлялось

процентное соответствие близости текста к конкретному классу. Лингвистические особенности всех текстов различались и были проанализированы.

Для автоматической классификации текстов деление собранного материала производилось посредством инструментов, которые выдавали более близкую, к слову, категорию для прибавления.

Из библиотеки keras были применены модели и layers. Keras — открытая нейросетевая библиотека, написанная на языке Python [4]. Целью Keras является оперативная работа с глубоким обучением. Собранные данные были приведены в исходную форму в массиве словосочетаний, а вслед затем и текстов. Для обеспечения очередности в перечне по схожести длины была применена pad_sequences из keras. По умолчанию это делается путем добавления 0 в начале каждой последовательности, пока каждая последовательность не будет иметь такую же длину, как и самая длинная последовательность. Для выявления категорий был применен алгоритм машинного обучения Multinomial Naive Bayes. Тренировка позволяет минимизировать издержки.

В рамках реализации нашего инструмента для систематизации по категориям мы сделали наш классификатор. Для этого были применены техники удаления стоп-слов, замены прописных букв на строчные и приведения текстов в исходную форму. Дальше была задействована функция Tokenizer из библиотеки keras. Tokenizer преобразовывал слова в очередности и переводил их в векторное представление как числовое. Дальше с поддержкой pad_sequences из библиотеки keras preprocessing мы возвращали очередности с интервалами и метками.

Для наибольшей сложности систематизации категорий полученные результаты из нашего классификатора по работе с входными данными мы запускали в функцию predict нашей модели, построенной на основе MultinomialNB (Multinomial Naive Bayes). Multinomial Naive bayes — один из вариантов классификаторов, построенных на алгоритме NB, который используется в основном в задачах обработки полиномиально распределенных данных, например, в классификации текста [5]. Для оптимизации рабочего процесса мы предварительно сохраняли подборки итога тренировки модели в формате h5 и загружали итог через функцию библиотеки keras load_model. В итоге получали результат, с поддержкой индексирования итога выводили категорию. Первые результаты категоризации показывают точность модели около 80% (табл. 1).

Таблица 1. Первый результат категоризации текстов на примере большого корпуса данных.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.83 | 0.8 | 2777 |
| 1 | 0.82 | 0.77 | 0.8 | 2763 |
| micro-avg | 0.8 | 0.8 | 0.8 | 5540 |
| macro-avg | 0.8 | 0.8 | 0.8 | 5540 |
| weighted-avg | 0.8 | 0.8 | 0.8 | 5540 |

Чтобы обрабатывать разные случаи ввода, мы можем собирать данные из различных источников и сфер деятельности, так повысится качество собранного набора данных, что даст положительный эффект в работе модели. В процессе внедрения модели мы увеличивали точность модели с использованием логистической регрессии, и результат точности увеличился до 84%. Метод будет хорошо работать на различных типах данных, включая иронию и сатиру, так как постоянно идет улучшение за счет количества и качества данных.

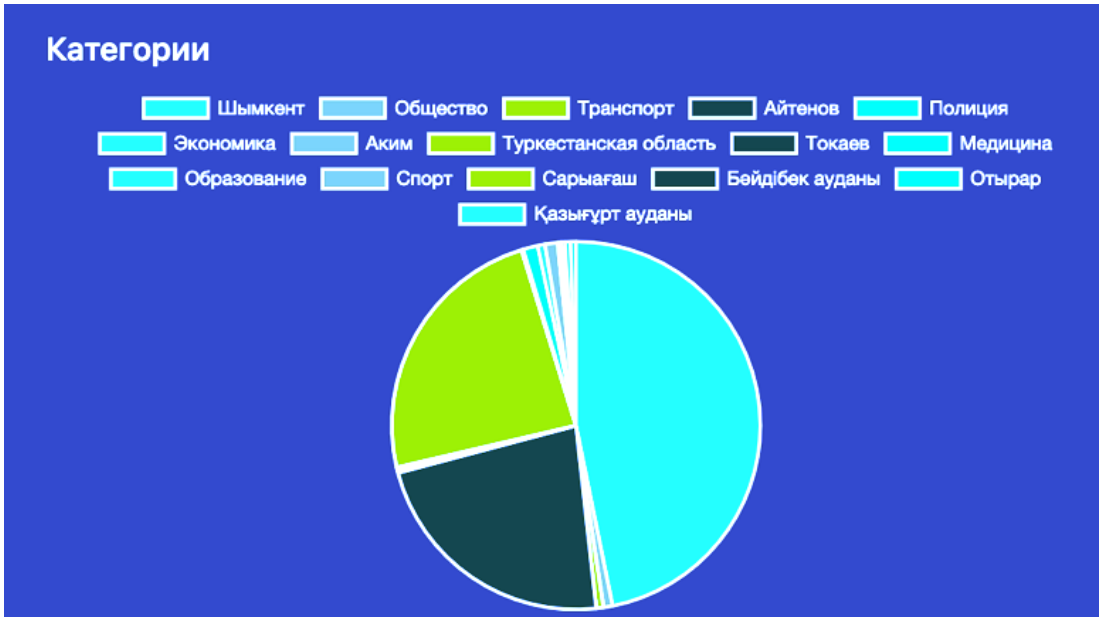



*Рисунок 2 - Результат категоризации текстов на примере большого корпуса данных после улучшений модели*

Как и все остальные системы, теоретически систему классификации, основанную на анализе текста, можно обманывать, если тексты приходят не в подготовленных шаблонах, но для этого есть процент погрешности. Несмотря на это, на практике система работает стабильно и показывает хорошие результаты по классификации текстов. Тексты, приходящие из информационных систем, а также из различных источников, классифицируются по категориям и дают возможность для последующей обработки и анализа естественного языка. Эта работа основана на гипотезе, что тексты являются неструктурированными и собираются из разных источников, но есть процент погрешности выявления категорий в зависимости от самой области контекста. Решением данной проблемы является увеличение корпуса данных, тренировка моделей и работа по чистке данных и выявлению ключевых параметров.

**Заключение**

В заключении нужно отметить, что целью классификации текстов по их признакам является обеспечение целостного понимания и обработки естественных языков. Классифицировать текст можно только сравнивая данные, получаемые от единиц текста, которые являются частью контекста. Так как общий контекст играет ключевую роль в классификации текста, необходимо учесть особенности языкового корпуса, токенизацию слов и приведение в изначальную форму. Для того, чтобы не было отклонений в точности результатов, необходимо работать над чистотой корпуса и уделять существенное внимание проработке модели и выстраиванию параметров. Таким образом данная модель будет иметь процент погрешности в расчетах, но будет постоянно самообучаться и давать результаты лучше с каждым разом, так как корпус текста нужно будет увеличивать и уделять внимание чистке.

СПИСОК ЛИТЕРАТУРЫ

1. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных, НИУ ВШЭ, Москва, 2017, 124-128 с.

2. Рафанов С.М. К проблеме классификации текстов в машинном переводе, КРСУ, Москва, 2013, 36-42 с.
3. Попова З.Д., Когнитивная Лингвистика, Федеральное агентство по образованию Воронежский Государственный Университет, Воронеж, 2007, 7-12 с.
4. Ketkar, Nikhil. (2017). Introduction to Keras. 10.1007/978-1-4842-2766-4_7.
5. Xu, Shuo & Li, Yan & Zheng, Wang. (2017). Bayesian Multinomial Naïve Bayes Classifier to Text Classification. 347-352. 10.1007/978-981-10-5041-1_57.

REFERENCES

1. Bol'shakova E.I., Voroncov K.V., Efremova N.E., Klyshinskij E.S., Lukashevich N.V., Sapin A.S. *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannyh* [Automatic Natural Language Processing and Data Analysis] NIU VSHE, Moscow, 2017, 124-128 c.
2. Rafanov S.M. K probleme klassifikacii tekstov v mashinnom perevode [the problem of text classification in machine translation], KRSU, Moscow, 2013, 36-42 c.
3. Popova Z.D., Kognitivnaya Lingvistika [Cognitive Linguistics], Federal'noe agentstvo po obrazovaniyu Voronezhskij Gosudarstvennyj Universitet, Voronezh, 2007, 7-12 c.
4. Ketkar, Nikhil. (2017). Introduction to Keras. 10.1007/978-1-4842-2766-4_7.
5. Xu, Shuo & Li, Yan & Zheng, Wang. (2017). Bayesian Multinomial Naïve Bayes Classifier to Text Classification. 347-352. 10.1007/978-981-10-5041-1_57.

**Жунусов Д.О., Алиаскаров С.Ж.**

**Машинналық оқыту алгоритмдері негізінде мәтіндер классификациясының әдісі**

**Аңдатпа.** Бұл жұмыста мәтінді жіктеу әдісін есептеу және талдау әдістері сипатталған. Сонымен қатар, табиғи тілді өңдеуге арналған машиналық оқыту алгоритмдерін қолданудың негізгі параметрлері анықталды. Бұл жұмыс үшін машиналық оқыту алгоритмдерінің кейбір енгізілімдері талданды және көрсетілді. Мәтіндерді талдау және санаттарға бөлу жүйесін құру үшін мәтіндік деректерді өңдеу алгоритмдері, сонымен қатар табиғи тілді өңдеу жүйесін зерттеу нәтижелері қолданылды.

**Түйінді сөздер:** өңдеу, табиғи тілдер, жіктеу, талдау, санаттарға бөлу, тану, түсіну.

**Zhunissov D.O., Aliaskarov S.Zh.**

**Method for text classification based on machine learning algorithms**

**Abstract.** This work describes the methods of calculating and analyzing the text classification and identifies the main parameters of applying machine learning algorithms for natural language processing. The authors analyzed some implementations of machine learning algorithms, applied such algorithms for processing text data to develop a system for the analysis and categorization of texts and demonstrated the results of their study of a natural language processing system.

**Keywords:** processing, natural languages, classification, analysis, categorization, recognition, understanding.

**Авторлар туралы мәлімет:**

**Алиаскаров Серик Женисханович**, «Ақпараттық жүйелер» кафедрасының докторанты, Халықаралық ақпараттық технологиялар университеті.

**Жунусов Дархан Ондасынулы**, «Ақпараттық жүйелер» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Алиаскаров Серик Женисханович**, докторант кафедры «Информационные системы», Международный университет информационных технологий.

**Жунусов Дархан Ондасынулы**, магистрант кафедры «Информационные системы», Международный университет информационных технологий.

**About the authors:**

**Serik Zh. Aliaskarov**, doctoral student, Department of Information Systems, International Information Technology University.

**Darkhan O. Zhunussov**, master student, Department of Information Systems, International Information Technology University.

# ЦИФРОВЫЕ ТЕХНОЛОГИИ В ЭКОНОМИКЕ И МЕНЕДЖМЕНТЕ

УДК 004.45, 005.6

**Алимжанова Л.М.[1], Панарина А.В.[2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан

## ВНЕДРЕНИЕ СЕРВИСНОЙ СИСТЕМЫ IT-АУТСОРСИНГА

**Аннотация.** Производительность многих организаций зависит от их способности эффективно использовать свои системы ИТ-услуг. Это подразумевает необходимость оптимизации отношений между поставщиком ИТ-услуг и сотрудниками организации. Для управления услугами используются так называемые ITSM-системы. Но выбрать из существующих или разработать новую систему — это только «верхушка айсберга». Успешно внедрить систему и сделать так, чтобы ей пользовались — одна из ключевых задач и важнейшее условие жизни самой системы.

В данном исследовании показано, как процесс внедрения системы ИТ-услуг «Asista» одной из аутсорсинговых компаний Казахстана анализируется и модернизируется на основе расчетов эффективности. Расчеты произведены посредством применения методик бережливого производства.

**Ключевые слова**: сервис, сервисная система, аутсорсинг, IT-аутсорсинг, бережливое производство, Information Technology Infrastructure Library, Information Technology Service Management.

### Введение

Надежная работа всей ИТ-инфраструктуры большинства компаний оказывает существенное влияние на эффективную и успешную реализацию бизнес-стратегий компании. Непрерывное комплексное развитие и усложнение структуры бизнес-процессов приводит к значительному повышению сложности управленческих задач и ресурсов, задействованных для их решения. Это также касается ИТ-услуг предприятий [1].

Передовой опыт организации IT-подразделений связан с концепцией, что основные бизнес-процессы современного предприятия, а также техническое и программное обеспечение, процессы жизненного цикла информационных систем и процессы взаимодействия с различными составляющими IT-инфраструктуры, в том числе конечными пользователями, должны рассматриваться как одно целое. Приведенную концепцию реализует ITSM (управление IT-услугами, англ. Information Technology Service Management) — подход, при котором IT-служба является полноправным элементом бизнес-структуры, а не вспомогательной службой [2].

Передовой опыт организации работы IT-подразделений как полноправного члена бизнес-процессов собран в ITIL (IT Infrastructure Library) — библиотеке инфраструктуры информационных технологий. Она имеет типовые шаблоны и модели, а также методологии, которые необходимо внедрить для эффективной организации работы не только IT-служб предприятия, но и всех зависимых от IT компонентов.

Преимущества применения ITIL и ITSM для IT-подразделений заключаются в возможности организовать эффективный механизм поддержки и сопровождения информационных систем предприятия, уменьшить количество сбоев при предоставлении необходимых сервисов, контролировать IT-инфраструктуру при изменениях, оценивать эффективность и работоспособность IT-служб, оптимизировать работу IT-служб в целом [3].

Существует целый класс программного обеспечения, который относят к ITSM-системам, например, такие системы выпускает корпорация BMC. IT-администраторы по всему миру решают схожие задачи. В некоторых случаях верно использовать лучшие практики и уже готовые системы, а в некоторых вернее будет написать свою уникальную систему, которая бы идеально подходила под все специфические задачи компании. Казахстанская компания, предоставляющая IT-аутсорсинг выбрала второй подход.

После написания сервисной системы «Asista» открытым остался вопрос: «Как эффективно внедрить данную систему среди нескольких тысяч клиентов компании?». На помощь пришла концепция бережливого производства (Lean Production), объединившая различные методы повышения эффективности процессов.

**Расчет эффективности процесса внедрения сервисной системы IT-аутсорсинга посредством применения технологий бережливого производства**

На сегодняшний день технология бережливого производства отвечает главному запросу всех предприятий — повышению эффективности процессов при условии ограниченности ресурсов. «Бережливое производство» подразумевает не просто краткосрочные меры по сокращению затрат на персонал, аренду и содержание складских помещений и другое, а в первую очередь — оптимизацию бизнес-процессов с целью исключения лишних и ненужных функций и процедур, создающих только дополнительную работу, но не создающих дополнительной ценности [4].

Бережливое производство направлено на устранение потерь во всех сферах деятельности предприятия, начиная с построения взаимоотношений с потребителями: в проектировании продукции, выстраивании цепей обеспечения, производственном менеджменте, транспортных и иных операциях. Целью такого производства является достижение наименьших трудозатрат, минимальных сроков создания продукции, гарантированной поставки продукции покупателю в назначенное время, высокого качества при наименьшей стоимости. При этом под потерями необходимо понимать любое действие, при выполнении которого используются ресурсы, но не создается ценность для потребителя продукции (клиента) [5].

Бережливое производство нацелено на сокращение потерь (именно потерь, а не затрат). Выделяют несколько видов потерь: ожидание, дефекты и переделки, передвижения, запасы, излишняя обработка, перемещение материалов, перепроизводство. Эти потери увеличивают издержки производства, не добавляя потребительской ценности, действительно необходимой заказчику [6].

Основной процесс, который мы будем анализировать — это процесс внедрения ИТ-системы «Asista». На первом этапе работы осуществляется картирование производственных процессов с целью визуализации порядка оказания услуг путем построения карты потока (согласно терминологии бережливого производства). Нотацию будем использовать BPMN (англ. Business Process Model and Notation), так как с ее помощью очень хорошо просматривается хронометраж процесса.

Обозначим основные роли: Сотрудник, Клиент, CRM-система и вспомогательная роль — Документы. Также выпишем основные объекты потока управления: события (в нашем случае — начало и окончание процесса), действия (например: «выезжает к клиенту») и логические операторы (рис. 1).

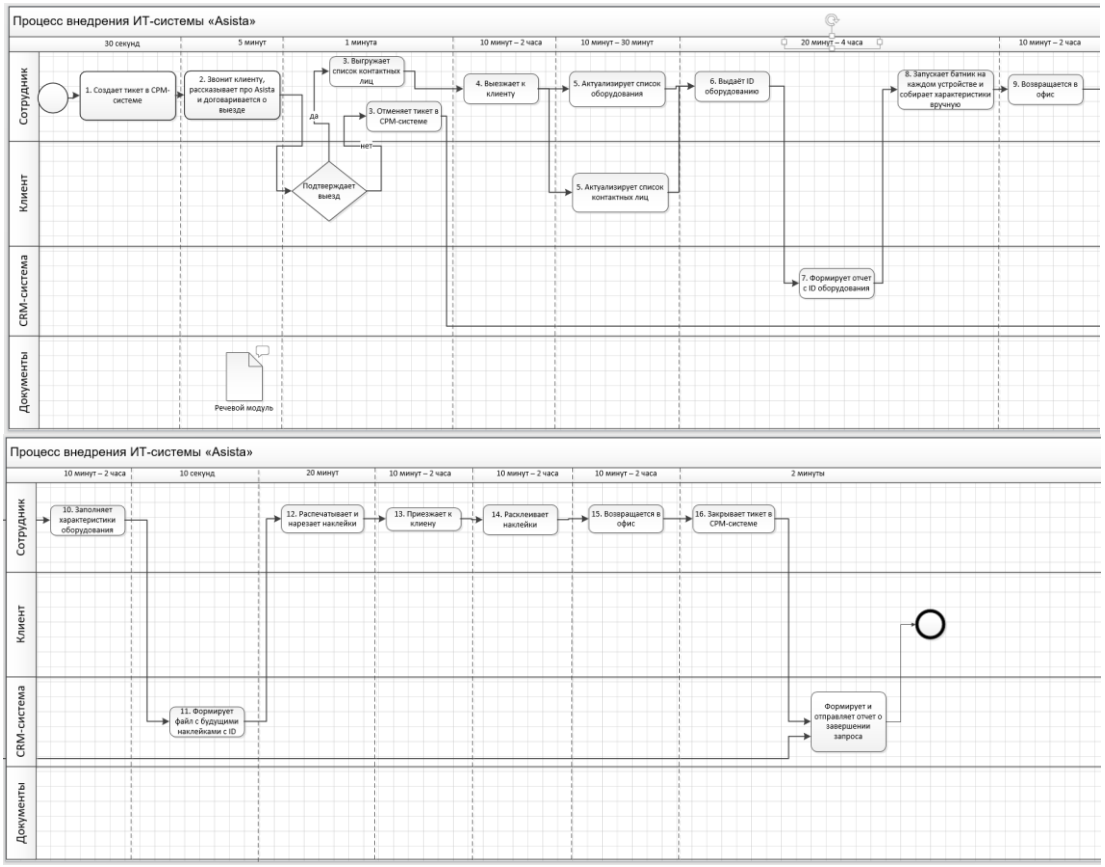



*Рисунок 1 - Текущее состояние процесса внедрения ИТ-системы «Asista»*

Эффективность потока рассчитаем по формуле:

$$Кэф = \frac{ΣВСЦ}{ΣВСЦ+ΣТп} * 100\% \quad (1)$$

где: Кэф — коэффициент эффективности потока; ΣВСЦ — суммарное время создания ценности на всех операциях потока; ΣТп — суммарное время потерь в потоке.

При описании потока создания ценности, включающего все этапы предоставления услуги потребителю или производства продукта, все действия делятся на три категории:

1) действия, которые создают ценность для потребителя (это ключевые действия, во время таких действий «меняется» сам продукт или услуга);

2) действия, которые, с точки зрения потребителя, не создают ценности, но которые нельзя убрать из процессов (муда первого рода);

3) действия, которые не создают ценности, и поэтому могут быть немедленно исключены из потока (муда второго рода) (лишние движения, лишняя транспортировка) [7].

В нашем случае ценность приносят такие действия, как: «запускает бат-файл на каждом устройстве» (момент, когда создается ярлык на рабочем столе клиента) и «расклеивает наклейки» (момент, когда видоизменяется устройство: на нём появляется наклейка с идентификатором).

$$Кэф = \frac{360\ минут}{360\ минут + 538{,}65\ минут} * 100\% = \frac{360\ минут}{898{,}65\ минут} * 100\% = 40\%$$

Таким образом, наш процесс эффективен на 40%, что не составляет даже половины от затраченных на него ресурсов. После проведенного анализа стало понятно, что большая

часть времени уходит на передвижения — приезд к клиенту и возвращение в офис, и так дважды.

Потери при передвижении — это ненужные перемещения материалов, готовых изделий, персонала и оборудования, которые не добавляют ценности процессу.

Делаем вывод о необходимости сократить время на передвижения. Для этого сокращаем количество выездов к клиенту с двух до одного. Сотрудник возвращается в офис только для того, чтобы распечатать наклейки. Для того, чтобы сократить количество выездов, было предложено два варианта:

1. При себе иметь печатающее устройство, с помощью которого можно распечатать наклейки на оборудование с идентификационным номером;
2. Распечатывать наклейки с идентификационными номерами с запасом до выезда к клиенту.

После анализа рынка печатающих устройств была выявлена нецелесообразность покупки. Большинство портативных устройств, предназначенных для работы с самоклеящейся бумагой, печатают только в черно-белом формате. Для компании важно наличие цветной распечатки. Поэтому процесс выдачи ID устройствам был полностью пересмотрен, что отразилось при составлении карты будущего состояния потока (рис.2).

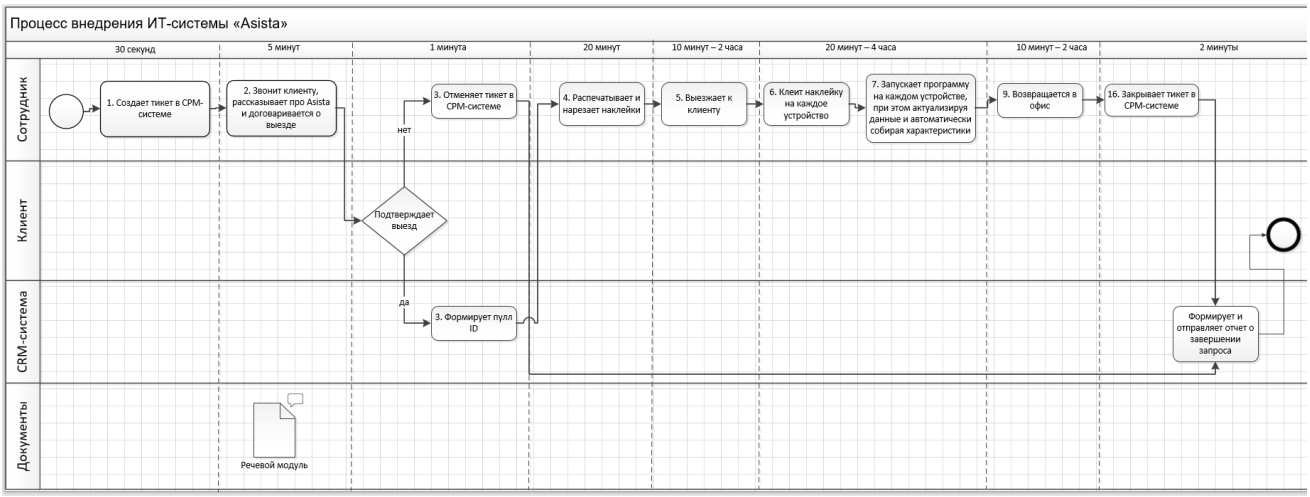



*Рисунок.2 - Будущее состояние процесса внедрения ИТ-системы «Asista»*

Рассчитаем коэффициент эффективности потока:

$$\text{Кэф} = \frac{240 \text{ минут}}{240 \text{ минут} + 268{,}5 \text{ минут}} * 100\% = \frac{240 \text{ минут}}{508{,}5 \text{ минут}} * 100\% = 48\%$$

Мы получаем увеличение эффективности на 8%. Это не так много в рамках данного процесса, но общее время цикла уменьшилось на целых 45%.

$$\text{Время цикла} = \Sigma\text{ВСЦ} + \Sigma\text{Тп} \quad (2)$$

Время цикла (текущее состояние) = 898,65 минут

Время цикла (будущее состояние) = 508,5 минут

Таким образом, использование инструментов бережливого производства снизит уровень временных потерь на внедрение сервисной системы IT-аутсорсинга путем исключения потерь второго уровня.

**Заключение**

Результаты проведенного посредством применения методик бережливого производства исследования заключаются в подтверждении необходимости улучшения существующего процесса внедрения сервисной системы IT-аутсорсинга.

Выявлена ценность знаний о данных методиках и инструментах современного бережливого производства, которые могут быть использованы для повышения эффективности не только проанализированного процесса, но и всей производственной системы предприятия. Все это даст возможность значительно повысить эффективность производственных процессов и улучшить организацию процесса оказания услуг.

СПИСОК ЛИТЕРАТУРЫ

1. Суровцев, А. Модели ITSM и системы ITIL как основа систем управления. Центральный научный вестник, (Номер 18 (35)), issn: 2499-9989, 2017. — С.10-11
2. Коротков Е.А. Роль информационных технологий в развитии предприятий // Сб. «Стратегические инициативы социально-экономического развития хозяйствующих субъектов региона в условиях внешних ограничений». Материалы международной научно-практической конференции, организованной совместно с администрацией ОЭЗ «ППТ «Липецк». — Елец: Елецкий государственный университет им. И.А. Бунина, 2017. — С. 136-138
3. Черпаков И.В. Обеспечение конкурентоспособности программного продукта в рамках разработки архитектурно значимых требований // Сб. «Стратегические инициативы социально-экономического развития хозяйствующих субъектов региона в условиях внешних ограничений». Материалы международной научно-практической конференции, организованной совместно с администрацией ОЭЗ «ППТ «Липецк». — Елец: Елецкий государственный университет им. И.А. Бунина, 2017. — С.344-349.
4. Фейгенсон Н.Б., Мацкевич И.С., Липецкая М.С. Бережливое производство и системы менеджмента качества: серия докладов (зеленых книг) в рамках проекта «Промышленный и технологический форсайт Российской Федерации» — СПб., 2012. — С. 9-10
5. Еманаков И.В., Овчинников С.А., Грудзинский П.В. Выявление и анализ потерь при внедрении бережливого производства на промышленных предприятиях. Качество и жизнь, 2014. — С. 56-58.
6. Давыдова, Н.С. Бережливое Производство: Монография. Изд-Во Института Экономики и Управления, ГОУВПО «УдГУ», 2012.
7. Джонс, Д., Вумек Д. Бережливое Производство: Как Избавиться От Потерь и Добиться Процветания. Альпина Паблишер, 2018.

REFERENCES

1. Surovtsev, A. ITSM models and ITIL systems as the basis of control systems. Central Scientific Bulletin, (No. 18 (35)), issn: 2499-9989, 2017. – pp.10-11
2. Korotkov E.A. The role of information technology in the development of enterprises // Coll. "Strategic initiatives for the socio-economic development of economic entities in the region in the context of external constraints." Materials of the international scientific and practical conference, organized jointly with the administration of the SEZ "PPT" Lipetsk ". - Yelets: Yelets State University named after I.A. Bunin, 2017. – pp. 136-138
3. Cherpakov I.V. Ensuring the competitiveness of a software product in the framework of the development of architecturally significant requirements // Coll. "Strategic initiatives for the socio-economic development of economic entities in the region in the context of external constraints." Materials of the international scientific-practical conference, organized jointly with the administration of the SEZ "PPT" Lipetsk ". - Yelets: Yelets State University named after I.A. Bunin, 2017 . – pp. 344-349.

4. Feygenson N.B., Matskevich I.S., Lipetskaya M.S. Lean manufacturing and quality management systems: a series of reports (green books) within the framework of the project "Industrial and technological foresight of the Russian Federation" - St. Petersburg, 2012. – pp. 9-10
5. Emanakov I.V., Ovchinnikov S.A., Grudzinsky P.V. Identification and analysis of losses in the implementation of lean manufacturing in industrial enterprises. Quality and Life, 2014. – pp. 56-58.
6. Davydova, NS Lean Manufacturing: Monograph. Publishing House of the Institute of Economics and Management, GOUVPO "UdSU", 2012.
7. Jones, D., Wumek D. Lean Manufacturing: How to Get Rid of Waste and Prosperity. Alpina Publisher, 2018.

**Алимжанова Л.М., Панарина А.В.**

**IT-аутсорсингтің сервистік жүйесін енгізу**

**Аңдатпа.** Көптеген ұйымдардың өнімділігі олардың IT-қызмет жүйелерін тиімді пайдалану қабілетіне байланысты болады. Бұл провайдер мен ұйым қызметкерлері арасындағы қатынастарды оңтайландыру қажеттілігін білдіреді. Қызметтерді басқару үшін ITSM деп аталатын жүйелер қолданылады. Бірақ бар жүйені таңдау немесе жаңа жүйені жасау – бұл «айсбергтің ұшы» ғана. Жүйенің негізі міндеттері мен тіршілік жағдайлары – аталған жүйені сәтті енгізу және пайдалану.

Берілген зерттеуде Қазақстанның аутсорсингтік компанияларының бірі «Asista» IT-қызметтер жүйесін енгізу процесі тиімді есептеу негізінде жұмыс ісі талданып, қайта жаңғыртылатыны көрсетілген. Үнемді өндіріс әдістерін қолдану арқылы есептеулер жүргізілді.

**Түйінді сөздер**: сервис, қызмет көрсету жүйесі, аутсорсинг, IT-аутсорсинг, үнемді өндіріс, Information Technology Infrastructure Library, Information Technology Service Management.

**Alimzhanova L.M., Panarina A.V.**

**Implementation of an IT outsourcing service system**

**Abstract.** The performance of many organizations depends on their ability to effectively use their IT service systems. This implies the need to optimize the relationship between the IT service provider and the people in the organization. The so-called ITSM systems are used to manage the services. But choosing from the existing ones or developing a new system is only the “tip of the iceberg”. Successfully implementing such a system and making sure that it is used effectively is one of the key tasks and living conditions of the system itself.

This study focuses on the analysis and modernization of the “IT services system of Asista”, one of the outsourcing companies in Kazakhstan, based on efficiency calculations made using lean production techniques.

**Keywords:** service, service system, outsourcing, IT outsourcing, lean production, Information Technology Infrastructure Library, Information Technology Service Management.

**Авторлар туралы мәлімет:**

**Алимжанова Лаура Муратбековна**, т.ғ.к., «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті.

**Панарина Александра Владимировна**, «Ақпараттық жүйелер» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Алимжанова Лаура Муратбековна**, к.т.н., ассоциированный профессор кафедры «Информационные системы», Международный университет информационных технологий.

**Панарина Александра Владимировна**, магистрант кафедры «Информационные системы», Международный университет информационных технологий.

**About the authors:**

**Laura M. Alimzhanova,** Cand. of Sc. (Technology), Associate Professor, Department of Information Systems, International Information Technology University.

**Alexandra V. Panarina,** master student, Department of Information Systems, International Information Technology University.

УДК 004.45, 005.6

**Жұмабай Р.Ж.[1], Алимжанова Л.М.[2]**

[1,2] Международный университет информационных технологий, Алматы, Казахстан

# УПРАВЛЕНИЕ ПРОЦЕССАМИ РАБОТЫ С ПОСТАВЩИКАМИ НА ОСНОВЕ ERP-СТАНДАРТОВ — ПОДХОД BPM

**Аннотация**. Эффективное внедрение системы ERP на пути от покупателя к поставщику может значительно снизить материальные затраты и количество времени на производственный процесс. Этот документ представит BPM (Управление бизнес-процессами) подход к работе с поставщиками на основе ERP-стандартов. Рабочие процессы определяются с использованием стандарта BPMN (нотация и модель бизнес-процессов), проектируются и моделируются с использованием инструмента BPM. Метрики собираются для управления процессами. В данной статье рассматривается важность корректной настройки процесса документооборота и правильного определения ответственных за каждую операцию с входящими и исходящими документами, а также выделяются ключевые аспекты работы с поставщиками при поддержке методологии и инструментов BPM.

**Ключевые слова:** Enterprise Resource Planning, планирование ресурсов предприятия, управление бизнес-процессами, работа с поставщиками, Business Process Management, моделирование бизнес-процессов, Business Process Management Notation.

## Введение

ERP — это информационная система для планирования и интеграции всех подсистем предприятия, включая финансы, закупки, производство, человеческие ресурсы и продажи. Основная функция ERP заключается в интеграции процедур межведомственного управления и информационных систем управления. ERP эффективно снижает стоимость цепочки поставок, сокращает время производства, улучшает качество продукции, обеспечивает лучшее обслуживание клиентов и уравновешивает прогнозируемый спрос и предложение. Система интеграции ERP включает в себя основные подразделения предприятия, которое нужно связать и интегрировать [1].

Выбор поставщика и распределение заказов являются наиболее значимыми проблемами для отдела закупок предприятия, и эти две проблемы могут быть эффективно интегрированы в систему ERP для оптимизации процесса закупки. Выбор подходящих поставщиков не только дает существенные выгоды для компании, но также повышает удовлетворенность клиентов. Проблема выбора поставщика является проблемой группового решения по многочисленным критериям, а также неопределенным и неточным данным. Производители должны взаимодействовать с поставщиками, чтобы максимизировать производительность и минимизировать общую стоимость выпускаемой продукции [2].

## Управление бизнес-процессами (BPM)

Процесс — это набор взаимосвязанных задач, которые вместе преобразуют входные данные в выходные данные, выполняемых человеком или машиной.

Управление бизнес-процессами (BPM) — это набор методов, инструментов и технологий, используемых для разработки, принятия, анализа и контроля операционных бизнес-процессов. BPM — это процессно-ориентированный подход для улучшения производительности, который объединяет информационные технологии с методологиями процессов и управления. BPM охватывает людей, системы, функции, предприятия, клиентов, поставщиков и партнеров и подразумевает взаимодействие деловых людей и

информационных технологий для стимулирования эффективных, гибких и прозрачных бизнес-процессов [3].

При внедрении BPM в организации, чтобы улучшить производственные процессы, мы должны использовать проектный подход.

Ядром развертывания проекта BPM является моделирование процесса, которое должно учитывать и определять следующие элементы: объем процесса, запуск процесса, процесс деятельности и их взаимосвязь, номера процессов (измеримые цифры деятельности). Необходимые ресурсы и их распределение (продолжительность процесса, количество сотрудников, максимальная нагрузка, ключевые показатели эффективности процесса (KPI), завершение процесса и технологические связи с другими процессами (системная перспектива всех процессов в компании) должны быть задокументированы до начала внедрения автоматизации.

Анализ «AS-IS» («как есть») бизнес-процесса, включающий в себя обозначение цели процесса, графическое представление процесса (с использованием BPMN-нотации и модели бизнес-процессов), описание деятельности, KPI (ключевые показатели эффективности) и метрики, а также ссылки на другие процессы, должен быть выполнен в начале работы [4].

На основе собранных данных будет смоделирован процесс улучшения. Финальный процесс называется «Процесс TO BE» («как должно быть») — это корректировка цели. Предлагаемая модель процесса должна быть лучше исходной. Некоторые инструменты BPM имеют функцию моделирования процесса [4].

Модель бизнес-процесса постоянно совершенствуется и до окончательного принятия может претерпеть несколько изменений.

**Карта процессов системы**

Карта процессов — это инструмент для планирования и управления, с помощью которого удобно описывать течение работы [5]. Она включает в себя процессы управления, основные процессы, вспомогательные процессы, а также клиента и предлагаемые продукты. Именно основные бизнес-процессы (в данном случае — планирование и размещение заказа, расчет потребностей в товаре, анализ предложений поставщиков, выбор поставщика, формирование поставки, заключение договора) и создают ценность продукта. Остальные процессы, которые не являются главными ценностями, относятся к вспомогательным (в данном случае — анализ поставщиков, обеспечение безопасности документов, IT-обеспечение, обеспечение коммуникации с поставщиками, подготовка менеджеров). Ценность продукта клиент определяет сам через определенные требования. Ниже представлена карта процессов (рис. 1).

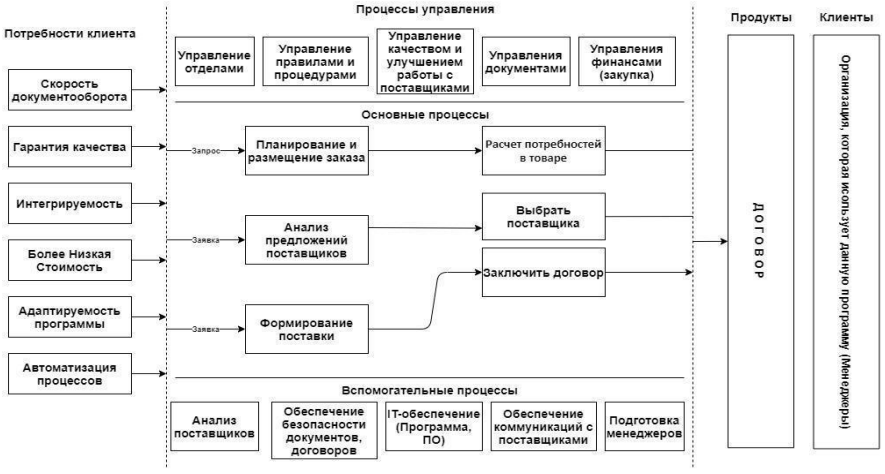



*Рисунок 1 - Карта процессов*

Далее продемонстрированы подпроцессы бизнес-процесса «Работа с поставщиками» (табл. 1).

Таблица 1 – Подпроцессы бизнес-процесса «Работа с поставщиками»

| | Подпроцесс | Ответственный | Входящие документы | Исходящие документы |
|---|---|---|---|---|
| 1а | Изучение внутренней статистики продаж | Специалист отдела планирования | Отчет-таблица собственных продаж | Нет |
| 1б | Изучение внешней статистики продаж | Специалист отдела планирования | Отчет-таблица продаж внешних источников | Нет |
| 2 | Расчет потребностей в товаре | Специалист отдела планирования | Отчет-таблица собственных продаж; Отчет-таблица продаж внешних источников | Таблица потребностей в товаре |
| 3 | Планирование и размещение заказа | Специалист отдела планирования | Заявка | План закупок |
| 4 | Ввод в систему прайс-листов поставщиков | Специалист отдела закупок | Справочник цен поставщиков | Прайс-листы поставщиков |
| 5 | Анализ предложений поставщиков и действующих контрактов | Специалист отдела закупок | Справочник цен поставщиков; Существующие контракты | Список поставщиков |
| 6 | Выбор поставщиков | Специалист отдела закупок | Перечень поставщиков | Список поставщиков с расстановкой приоритетов |
| 7 | Формирование поставки | Специалист отдела закупок | Перечень поставщиков с расстановкой приоритетов; Таблица потребностей в товаре | График поставок |
| 8 | Направление заказа в отдел закупок | Специалист отдела логистики | Заказы, принятые поставщиком | Установленные заказы |
| 9 | Направление заказа поставщику | Специалист отдела закупок | Заказы, принятые поставщиком | Установленные заказы |
| 10 | Заключение договора | Специалист отдела закупок | Договор | Договор |

**Модель бизнес-процессов «AS IS»**

Создание модели существующей организации процесса «AS-IS» («как есть») позволяет определить неэффективные места существующего на момент моделирования процесса [4]. Ниже изображена модель бизнес-процесса «AS-IS» (рис. 2).

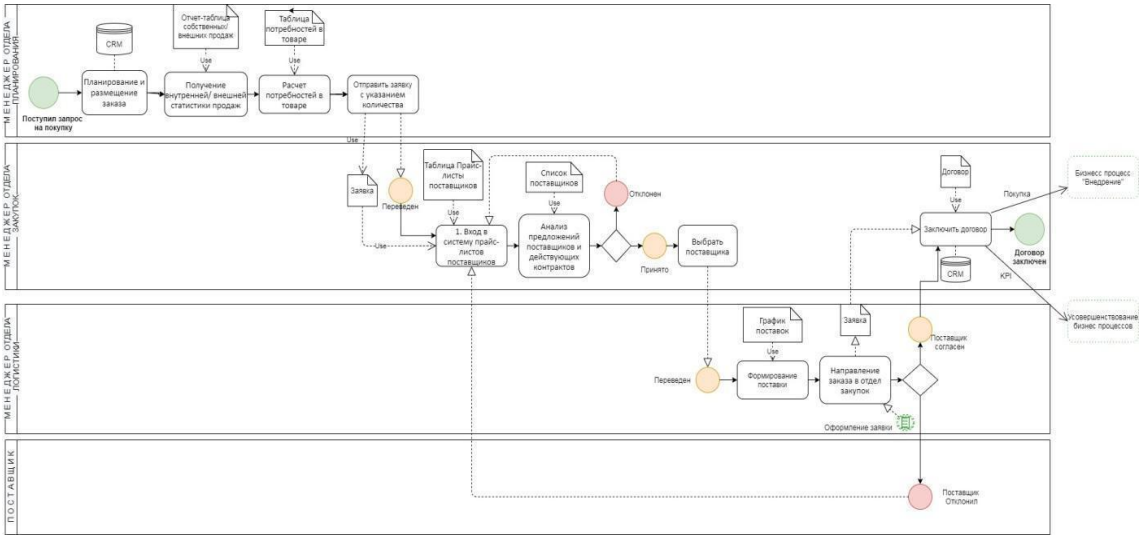



*Рисунок 2 - Модель «AS-IS»*

**Модель бизнес-процессов «TO BE»**

После того как бизнес-процессы описаны и смоделированы, мы приступаем к анализу бизнес-процессов, выявлению сложностей их реализации, а также к поиску путей решения этих проблем и повышению эффективности реализации процессов. Для этого используется качественный метод, при котором измеряются показатели эффективности: изучается потребность во входах и выходах и выявляются неиспользуемые выходы [4]. При помощи SWOT-анализа определяются необходимые изменения и улучшения, такие, как добавление таймеров и KPI. Ниже представлена модель бизнес-процесса «TO BE» (рис. 3).

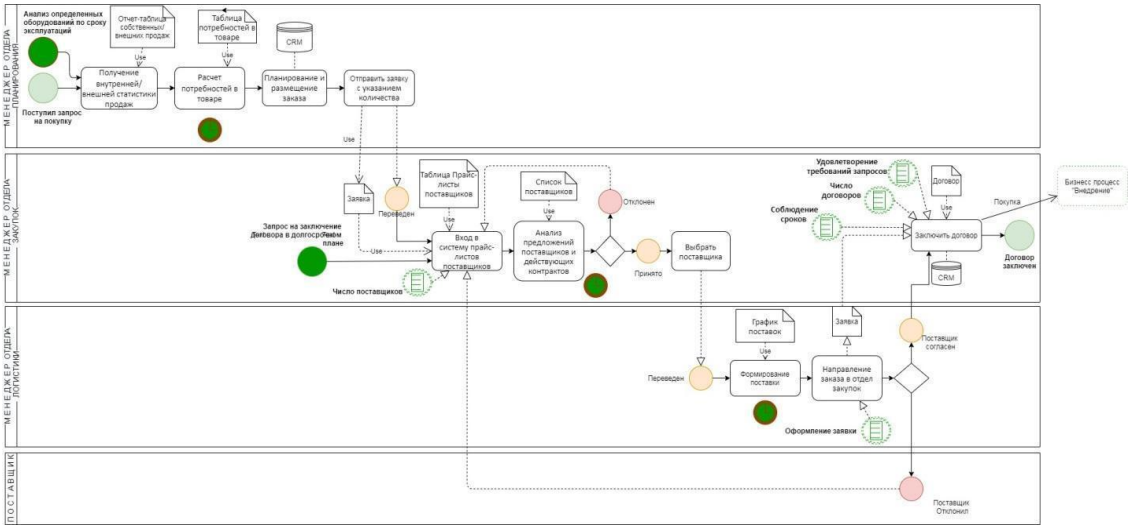



*Рисунок 3 - Модель «TO BE»*

**Заключение**

Проекты в сфере внедрения автоматизированных систем на современном предприятии давно перестали выглядеть чем-то своеобразным. Так же, как и любой другой проект, внедрение автоматизированной системы требует комплексной оценки как с точки зрения затрат, так и с точки зрения положительных эффектов, которые будут получены в результате. В настоящее время на рынке программных продуктов автоматизированных систем представлено огромное множество прикладных решений, позволяющих оптимизировать одни и те же бизнес-процессы.

В данной работе были проанализированы меры, необходимые для успешного продвижения ERP-системы на рынке.

Предлагаемый метод интуитивно понятен и прост в освоении, использовании и применении ко многим видам процессов, поскольку имеет графическое представление, облегчающее его восприятие членами команды.

СПИСОК ЛИТЕРАТУРЫ

1. Дэниел О'Лири. ERP системы. (2007) Системы планирования ресурсов предприятия: системы, жизненный цикл, электронная коммерция, риски (1-е изд., С. 112-156). Кембридж: Издательство Кембриджского университета. (дата обращения: 10.03.2020)
2. Tarantilis, C.D., K iranoudis, C.T. & Theodorakopoulos, N.D. (2008). Веб-система ERP для бизнес-услуг и управления цепочками поставок: приложение для планирования процессов в реальном мире. Operational Research, 187 (3), 1310-1326. DOI: 10.10161j. ejor.2006.09.015 (дата обращения: 10.03.2020)
3. Гаримелла, К., Лиз, М., и Уильямс, Б. (2008). Основы BPM для чайников (1-е изд., Т. 1, С. 30-40.). Хобокен, Нью-Джерси: Wiley Publishing (дата обращения: 20.03.2020)
4. Модель и нотация бизнес-процессов (BPMN) (2011, Январь 1)OMG [Электронный ресурс]. URL: http://www.omg.org/spec/BPMN/2.0/PDF (дата обращения: 20.05.2020)
5. Что такое карта процесса и как ее создать? Lucidchart [Электронный ресурс]. URL: https://www.lucidchart.com/pages/ru/карта-процесса (дата обращения: 10.05.2020)

REFERENCES

1. OLeary, D. E. (2007). *Enterprise resource planning systems: Systems, life cycle, electronic commerce, and risk* (1st ed., pp. 112-156.). Cambridge: Cambridge University Press. (accessed 10.03.2020)
2. Tarantilis, C.D., K iranoudis, C.T. & Theodorakopoulos, N.D. (2008). *Web-based ERP system for business services and supply chain management: a real-world process planning application.* Operational Research, 187 (3), 1310-1326. DOI: 10.10161j. ejor.2006.09.015. (accessed 10.03.2020)
3. Garimella, K., Lees, M., & Williams, B. (2008). *BPM basics for dummies* (1st ed., Vol. 1, pp. 30-40.). Hoboken, NJ: Wiley Publishing. (accessed 20.03.2020)
4. *Model i notacia biznes processov(BPMN)* [Model and notation of business processes] (2011, January 1)OMG [Электронный ресурс]. URL: http://www.omg.org/spec/BPMN/2.0/PDF (accessed: 20.05.2020)
5. *Chto takoe karta processa i kak ee sozdat'?* [What is a process map and how to create it?] Lucidchart [Electronic resource] URL: https://www.lucidchart.com/pages/ru/karta-protsessa (accessed 10.05.2020)

**Жұмабай Р.Ж., Алимжанова Л.М.**

**ERP стандарттарына негізделген жеткізушілермен жұмыс процесін басқару - BPM тәсілі**

**Аңдатпа.** Тапсырыс берушіден жеткізушіге дейінгі процесте ERP-ді тиімді енгізу айтарлықтай шығындар мен уақытты азайтуы мүмкін. Бұл құжат ERP стандарттарына негізделген жеткізушілермен жұмыс істеу процесінде BPM (Business Process Management) әдісін қолданады. Жұмыс процестері BPMN (Business Process Notation and Model) стандартының көмегімен анықталады және BPM құралының көмегімен құрастырылады және модельденеді. Метрика процесті басқару үшін жиналады. Бұл мақала жұмыс процесін жеңілдететін және BPM әдіснамалары мен құралдарының қолдауымен белгілі бір операцияларға кім жауап беретінін көрсететін және жеткізушілермен жұмыс процестерін жүзеге асырудың негізгі аспектілерін көрсететін таптырмас элемент ретінде енгізу және шығару құжаттарының маңыздылығын дәлелдейді.

**Түйінді сөздер:** Кәсіпорын ресурстарын жоспарлау, бизнес-процестерді басқару, жеткізушілермен жұмыс, бизнес-процестерді модельдеу, бизнес-процестерді басқару белгілері.

**Zhumabay R. Zh., Alimzhanova L.M.**

**Supplier process management based on ERP standards: the BPM approach**

**Abstract.** Effective customer-to-supplier ERP implementation can significantly reduce costs and time. This paper introduces a BPM (Business Process Management) approach to working with suppliers based on ERP standards. Workflows are defined using the BPMN (Business Process Notation and Model) standard and are designed and modeled using a BPM tool. Metrics are collected for process management. This article proves the importance of input and output documents as an indispensable element that will facilitate the workflow, show who is responsible for certain operations, and highlight the key aspects of implementing the process of working with suppliers, supported by BPM methodologies and tools.

**Keywords:** Enterprise Resource Planning, business process management, work with suppliers, business process modeling, Business Process Management Notation.

**Авторлар туралы мәлімет:**

**Алимжанова Лаура Муратовна,** т.ғ.к., «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті.

**Жұмабай Рауан Жұмабайқызы**, «Ақпараттық жүйелер» кафедрасының магистранты**,** Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Алимжанова Лаура Муратовна**, к.т.н., ассоциированный профессор кафедры «Информационные системы», Международный университет информационных технологий.

**Жұмабай Рауан Жұмабайқызы**, магистрант кафедры «Информационные системы», Международный университет информационных технологий.

**About the authors:**

**Laura M. Alimzhanova**, Cand. of Sc. (Technology), Associate Professor, Department of Information Systems, International Information Technology University.

**Rauan Zh. Zhumabay**, master student, Department of Information Systems, International Information Technology University.

УДК 33, 338,2, 338.24

**Berdykulova G.M.[1] Tolepbergenova D.A.[2]**

[1,2] International Information Technologies University, Almaty, Kazakhstan

# UNIVERSITY MANAGEMENT: CASE STUDY OF IITU

**Abstract.** The paper considers the university management practices based on the case study of IITU. It gives a review of the relevant literature to clarify the nature of educational and university management. The research was conducted using both the desk and field research methods. Study of the official documents allowed to reveal the practice and experience of the International Information Technology University in the management field. Surveys and interviews were conducted among teachers and students to understand how the university administration works. Finally, it presents a big picture of the IITU education management and its assessment through the eyes of its students and teachers**.**

**Key terms**: education management, university administration, university management, the digital transformation, the digital environment, the digital readiness, International Information Technology University.

## Introduction

According to the state program “Digital Kazakhstan”, educational management is one of the main directions in the country's development. A Kazakhstani university as an integral part of the social structure is influenced by the external factors one of which lately was the coronavirus pandemic which seriously affected the educational process. The situation requires a deep understanding of the causes and nature of the recent changes.

The paper explores the experience of university management, based on the case study of International Information Technology University, in order to find optimal ways of transforming its administrative system in the face of the current challenges. The original research of the IITU students and academic staff needs and problems helped to find out whether they are ready to work and study in the digital environment. Surveys of teachers and students on the IITU administration practices have revealed their shortcomings, problems of communication and readiness for distant learning and teaching.

## University and education management

In Kazakhstan, according to the Law of the Republic of Kazakhstan "On Education", educational organizations are managed in accordance with the legislation of the Republic of Kazakhstan, standard rules for the activities of educational organizations of the respective type and the charter of an educational organization operating under the principles of one-man management and collegiality [1].

The literature related to education management can be roughly divided into three areas:

-The first is the management of the educational process: training, education, or personality formation.

-The second is the management of educational organizations.

-The third direction is the management of education systems [2].

Professional knowledge in management determines the educational managers’ awareness of three different management tools:

-Organizations, management hierarchies, the main tool here is the impact on a person through motivation, planning, organization, control, incentives.

-Culture of management, embracing values, social norms and attitudes, behavioral features [3] developed and recognized by society, organization, group of people.

-Market, market relations based on the balance of interests of the seller and the buyer.

The effectiveness of management in the field of education consists in the well-coordinated functioning of educational organizations to provide quality education in a real economy through the implementation of the idea of lifelong education, expansion of the students' choice of educational paths in the conditions of reduced mobility. At present, there is a growing interest in improving management in the field of education under the State Program for the Development of Education of the Republic of Kazakhstan for 2011-2020. The main goal in the field of education management is the formation of a state-public education management system. The following objectives are set in pursuit of this goal:

-Improving management in education, including the introduction of corporate governance principles, the formation of a system of public-private partnership in education.

-Improving the system for monitoring the development of education, including national educational statistics, in compliance with the international requirements.

In this regard, the state educational policy of the Republic of Kazakhstan is a complex of interrelated measures covering changes in the structure, content and technologies of education and upbringing, organizational and legal forms of subjects of educational activity, financial and economic mechanisms, and the education management system per se.

**The administrative practices of International IT University**

The administrative structure of the university falls into four main sections as shown in the table below.

Table 1. The organizational structure of International IT University

| Section | Consistence |
|---|---|
| The first section | Vice Rector for Academic and Educational Activities, the Department of Academic Affairs, the Department of Postgraduate Education, the library, the Career Center, and the Scientific Methodological Council. |
| The second section | Vice-Rector for scientific and international activities, the Department of Science (NIR, SIS), the Dissertation Council (institutional body), the Department of International Cooperation and Academic Mobility. The work of the latter is mainly related to international relations, including student exchange, contacts with foreign partner universities, academic mobility programs, teacher internships abroad, contracting. |
| The third section | Vice-Rector for Innovation and Digitalization, Commercial Directorate, IITU Innovation Center, Department of Technical Support and IT Support, and Center for Educational Innovation and Smart Learning. These directorates and the IT Innovation Center work with qualified IT specialists to create robots, introduce new technologies, and improve the work of the IT Department. This department cooperates with the best specialists in the development of the university and the country's IT industry. |
| The fourth section | Military Department, Legal Department, Department of HR and Documentation, the Operations Directorate for economics and financial analysis, the Department of Accounting and Reporting, the Department of Marketing and PR, the Administrative Department. The fourth department is the organizational department of the university which performs work related to personnel, accounting, advertising, law. |

Compiled by the author on the basis of source [3]

The mission of IITU is the generation of knowledge and provision of training in the digital age. IITU Vision - University 4.0 [3].

Table 2. Strategic directions of IITU

| University brand positioning and promotion | |
|---|---|
| Purpose | Tasks |
| Entering the ranking of world universities | Participation in university rankings;<br>Digital marketing development;<br>Assistance in the spiritual modernization of youth;<br>Infrastructure development |
| Digital Excellence University | |
| Purpose | Tasks |
| Transformation of IITU into a digital university | Training for the digital age<br>Development of digital learning systems<br>Development and provision of digital services for stakeholders with the formation of a digital footprint<br>Gaining the position of a leading digital university |

Compiled by the author on the basis of source [3]

The IITU project team has been establishing contacts with the Higher School of Economics, the Moscow Institute of Physics and Technology and since October 2019 the Moscow School of Management Skolkovo to facilitate the process of its digital transformation. Thus, IITU began the digital transformation of the university driven by the introduction of platform, data storage and processing systems.

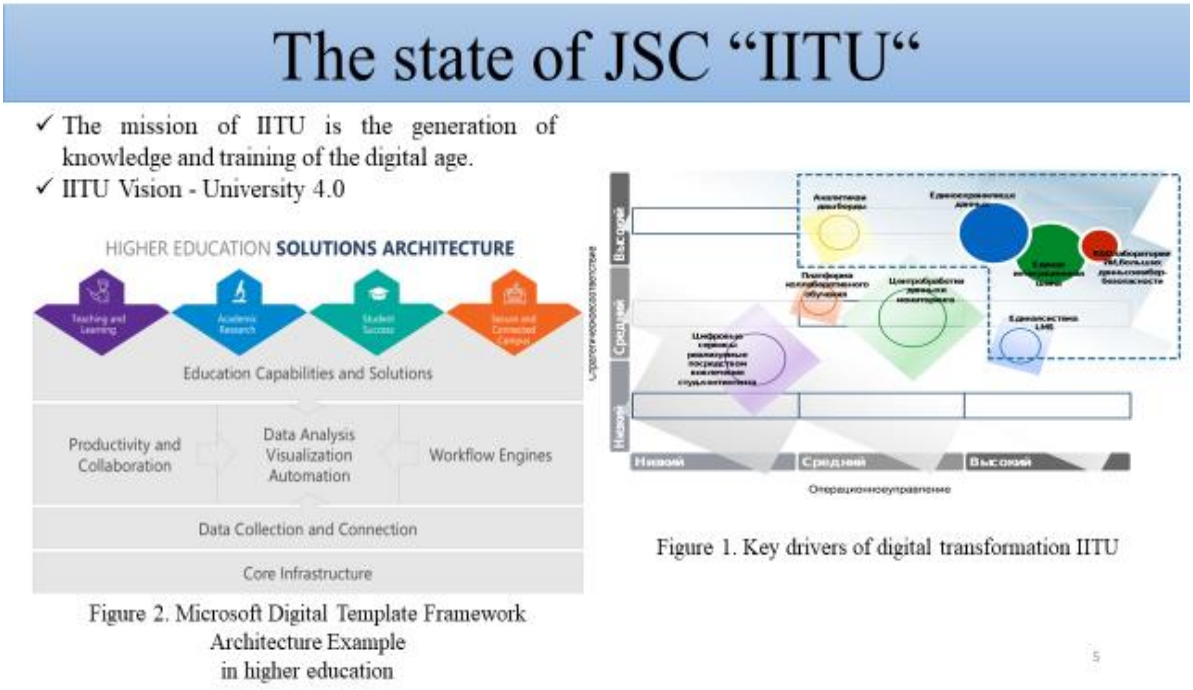



Figure 2. Microsoft Digital Template Framework Architecture Example in higher education

Figure 1. Key drivers of digital transformation IITU

*Figure 1 - The state of JSC "International IT University"*

**A survey and interview**

To understand how the IITU administration works, there was conducted a survey among its teachers and students. The main goal of the survey was to identify the problem areas in the work of the university management and to find out whether the university students and teachers are ready to work and study in the digital environment. The original research involves students' feedback on

IITU administration work, its shortcomings, problems of communication and readiness for working in the digital environment.

Table 3. IITU students' survey findings

| Interview with students | | | |
|---|---|---|---|
| Question's area | Number of students | Students category, % | Answers |
| Management of university | 38 | The 4th year- 81,6 | 57.9%, - satisfied, 42,1 % - dissatisfied. |
| Problems with the administration | | | The dean is always engaged; ten applications should be filed to retake an exam.<br>The administration staff don't want to solve the students' problems.<br>There is a need enable students to receive answers electronically and automatically. |
| Digitalizaton of the university | 38 | The 4th year- 81,6 | 80% - ready to study in a digital environment |

Presented below are the results of the student survey based on the original research.

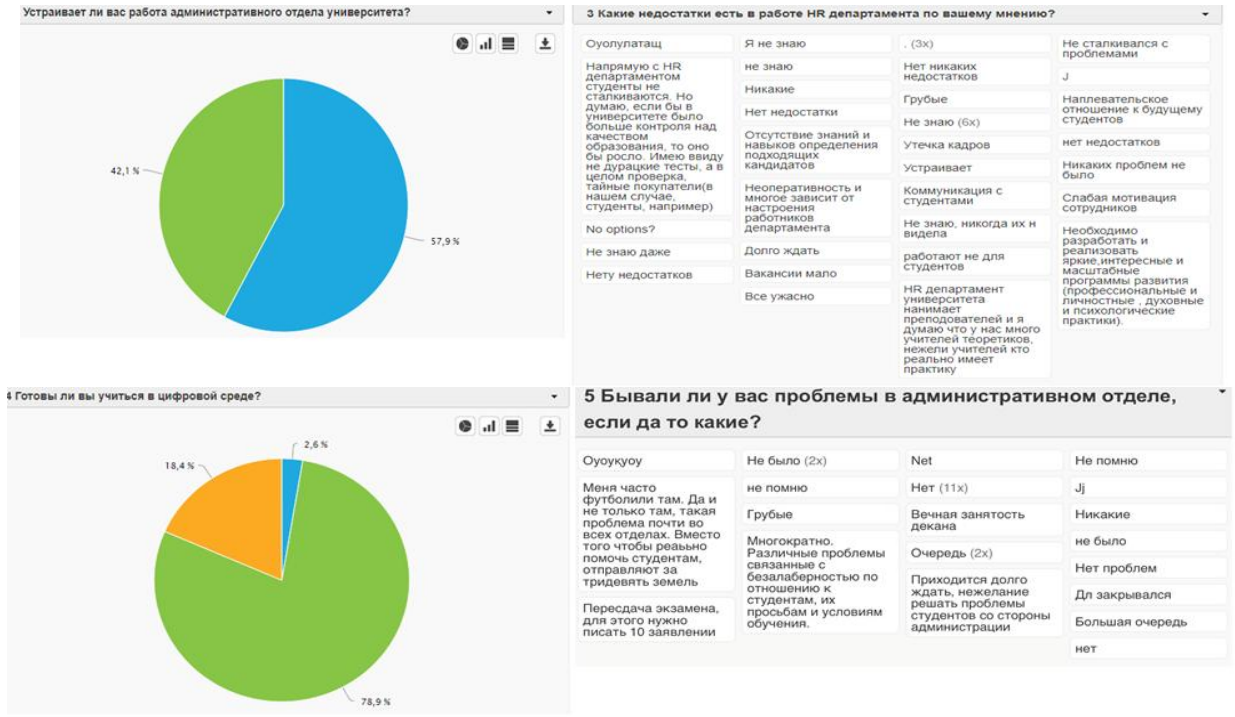



*Figure 2 - Survey of IITU students*

The results of a comprehensive assessment of digitalization of IITU show that joint work with Moscow universities has had a good effect on the digital transformation of management. The survey shows that more than 60% of the Department of Economics and Business and 80% of the university students are ready to work and study in the digital environment.

Teachers of the Department of Economics and Business participated in an online survey. Everyone has a computer knowledge of above average capacity, and 66.7% of teachers voted that they are ready to work in the digital environment.

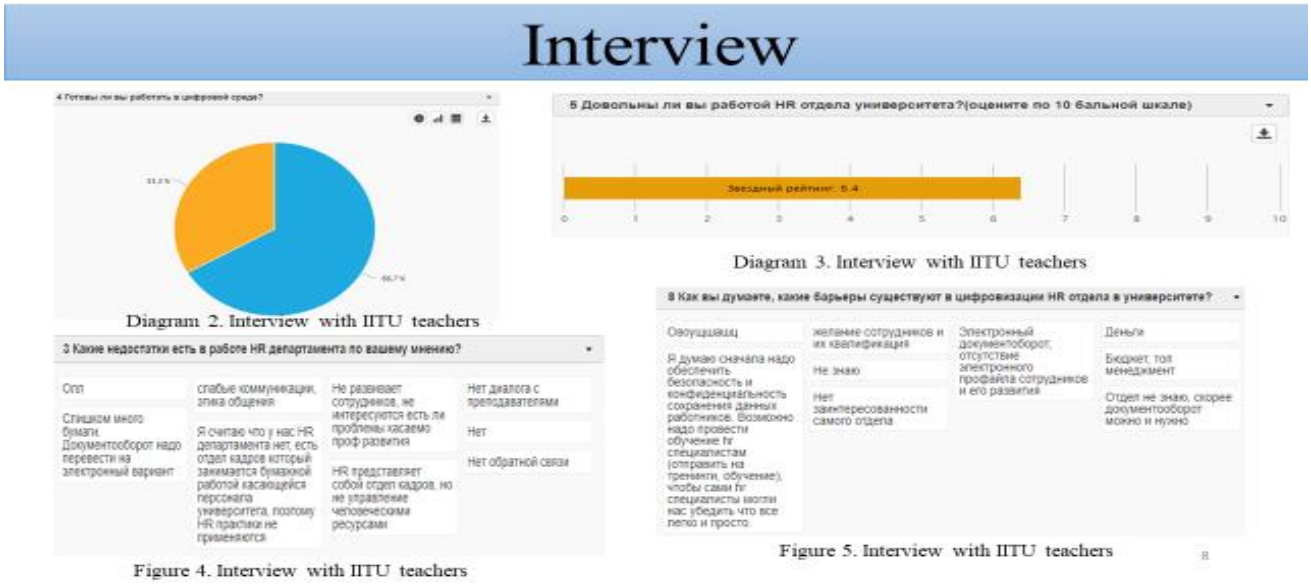



*Figure 3 - Interviews of the teaching staff of IITU Department of Economics and Business*

Teachers rated the work of the Personnel Department at 6.4 out of 10. In their opinion, the Personnel Department deals only with documentation, and this is only 1% of their work. This is because IITU does not have online platforms, for example, for storing documents or for the employee profiles.

**Conclusion**

Review of the relevant literature, examination of IITU administrative practices, surveys of and interviews with students and academic staff members on the respondents' satisfaction with the university administration and digitalization practices bring us to the conclusion that there are significant managerial flaws. Changes in the external economic environment in general and in the system of higher education in particular revealed the deficiency of management practices in the digital era, caused by the technological boom. Taking into account the provisions of the theory of post-industrial society and the new paradigm of management, it seems erroneous to introduce the principles of corporate governance in the formation of the educational management system at universities. Analysis of the organizational structure of IITU, the functional responsibilities of the highest management echelon, surveys and interviews with the stakeholders confirm the existence of a vertical hierarchy that reduces the possibility of effective university management. Yet, the digitalization processes at IITU are gaining momentum despite the problems caused by administrative barriers. Subsequent digitalization should be based on the theoretical provisions of change management, as the basis for the strategic development of the higher school and the university administration.

REFERENCES

1. The Law of the Republic of Kazakhstan dated 27 July, 2007 No. 319-III. [Электронный ресурс]. URL: https://iqaa.kz/images/Laws/Law_On_Education.eng.pdf (дата обращения 07.10.2020).
2. Shamsutdinova R.I., Baranova N.A. Menedgement v obrazovatelnoi deyatelnosti //Ekonomicheskie nauki. Marketing i Menedgement. [Electronic resource] URL: http://www.rusnauka.com/1_NIO_2013/Economics/6_123780.doc.htm
3. (дата обращения 07.10.2020).

4. Mission and Strategy. Values and principles. Organizational structure. JSC “International Information Technologies University”. [Электронный ресурс] URL: https://iitu.edu.kz/ru/(дата обращения 01.03.2021).
5. Bogdanova N.V., Puzankova L.V. Features: application of multimedia technologies in teaching computer science disciplines taking into account digital education // Informatics and Applied Mathematics: interuniversity collection of scientific papers. 2019.-.No. 25. – P.29-34.
6. Kalimullina O.V., Trotsenko I.V. Modern digital educational tools and digital competence: analysis of existing problems and trends // Open education. - 2018.Vol. 22. No. 3.-P. 61–73.

**Бердыкулова Г.М., Төлепбергенова Д.А.**
**Университетті басқару: ХАТУ практикасы**

**Аңдатпа.** Мақала Халықаралық ақпараттық технологиялар университетінің мысалында университетті басқару практикасына арналған. Білім беру және университетті басқару табиғаты туралы тиісті әдебиеттерге шолу жасалған. Әдістеме жұмыс үстеліне де, далалық зерттеулерге де негізделген. Халықаралық ақпараттық технологиялар университетінің менеджмент саласындағы тәжірибесі мен ресми құжаттарды зерттеу тәжірибесін анықтады. Университет әкімшілігінің жұмыс істеу процесін түсіну үшін оқытушылар мен студенттер арасында сауалнамалар мен сұхбаттар өткізілді. Соңында білім беруді басқарудың жалпы көрінісі қалыптасты және жалпы басқару жүйесі оқушылар мен мұғалімдердің көзімен бағаланды.

**Түйінді сөздер:** білім беру менеджменті, университет әкімшілігі, университет менеджменті, цифрлық трансформация, Халықаралық ақпараттық технологиялар университеті.

**Бердыкулова Г.М., Толепбергенова Д.А.**
**Менеджмент университета: практика МУИТ**

**Аннотация**. Статья посвящена практике управления университетом на примере Международного университета информационных технологий. Дается обзор соответствующей литературы, разъясняющей природу управления образованием и университетом. Методология основана как на кабинетных, так и на полевых исследованиях. Изучение официальных документов позволило выявить практику и опыт Международного университета информационных технологий в сфере менеджмента. Среди преподавателей и студентов были проведены опросы и интервью, чтобы понять, как работает администрация университета. Наконец, была установлена общая картина управления образованием и дана оценка общей системы управления университетом глазами студентов и преподавателей.

**Ключевые слова:** управление образованием, администрация вуза, управление университетом, цифровая трансформация, Международный Университет Информационных Технологий.

**Авторлар туралы мәлімет:**

**Бердыкулова Галия Мертаевна,** э.ғ.к., «Экономика және бизнес» кафедрасының профессоры, Халықаралық ақпараттық технологиялар университеті.

**Төлепбергенова Данагуль Алмасовна,** «Экономика және бизнес» кафедрасының студенті, Халықаралық ақпараттық технологиялар университеті.

**Сведения об авторах:**

**Бердыкулова Галия Мертаевна,** к.э.н., профессор кафедры «Экономика и Бизнес» Международный университет информационных технологий.

**Төлепбергенова Данагуль Алмасовна** студент кафедры «Экономика и Бизнес», Международный университет информационных технологий.

**About the authors:**

**Galiya M. Berdykulova,** Candidate of Economic Sciences, Professor, Department of Economics and Business, International Information Technology University.

**Danagul A. Tolepbergenova,** student, Department of Economics and Business, International Information Technology University.

УДК 659.3

**Omarova A.Sh.[1], Alimzhanova L.M.[2], Tashtamysheva A.E.[3]**

[1,2,3] International Information Technology University, Almaty, Kazakhstan

# RESEARCH AND DEVELOPMENT OF METHODS FOR THE TRANSITION OF TRADITIONAL MARKETING TO DIGITAL FORMAT

**Abstract.** The article presents the basic concept of the transition to digitalization, which is a new pivotal direction in the modern world, predetermining the competitiveness of states and companies, as well as the quality of citizens' life. Digital marketing will expand the capabilities of enterprises and the economy, strengthen all its systems and components. All this confirms the need to develop a methodology for the transition to multichannel digital marketing using the wide opportunities provided by the modern infrastructure of the digital space with adapted innovative technologies. Such technologies help promote more effectively a product or service on the market thanks to their clear focus on the target audience, focus group research, and the consumer reaction analysis followed by retargeting based on the developed customer profile. So, this article aims to conduct theoretical research and develop a practical methodology for the transition from standard to digital marketing approaches with the introduction of innovative marketing tools.

**Keywords:** traditional marketing, digital marketing, promotion, targeting, positioning, social networks.

## Introduction

The beginning of the 21st century was marked by economic growth and globalization, which lead to various kinds of crises. Economic and social problems have come together in a complex system of relations, the solution of which requires complex understanding, thinking and the introduction of new technologies. The development and spread of the Internet have significantly changed and continue to change the way people meet these challenges and the way they make decisions. Life smoothly flows online. Marketing of a company is a very active field of application of innovative digital technology thus making offline marketing less and less effective. In accordance with this, such a direction as digital marketing, or "Digital Marketing", has been born and is rapidly developing. Its goal is to find various innovative solutions that demonstrate efficiency and quick feedback between business, government agencies, social services, and society. Digitalization is a new pivotal development in the modern world, predetermining the competitiveness of states and companies, as well as the quality of their citizens' life. For instance, promotion of digital marketing will expand the capabilities of enterprises and the economy, strengthen all its systems and components.

## The main concept for the development of a project for the transition of traditional marketing to digital format

The term "digital marketing" goes back to the 90-ies. Many companies began to try to move away from the classics of promotion in marketing and find other recipes that could allow them to spend less and get more. These tasks could be solved by digital marketing because it allows you to reach both online and offline consumers who use tablets and mobile phones, play games, download applications. This way, the brand can reach a wider audience, not limited to the Internet. In addition to the above benefits, digital marketing can collect clear and detailed data. Almost all user actions in the digital environment are recorded by analytical systems. This allows you to draw accurate conclusions about the effectiveness of different promotion channels, as well as draw up an accurate portrait of the buyer. Digital marketing has a flexible approach, since it allows you to attract an offline audience to the online market, and vice versa. For example, using the QR code on the flyer,

you can direct the user to the site. And at the same time, thanks to the email newsletter, you can invite subscribers to a seminar or another offline event.

The main advantage of digital marketing is the ability to build communication with the target audience where they spend more time. Today it is online.

There are several tactics and tools that fall under the heading of digital marketing. This is the company's website itself and digital marketing channels - online promotion and customer acquisition channels: SEO, online advertising, email marketing, sales funnel, content marketing, teaser advertising, SMM, etc.

The goal of marketing has always been to meet with the target audience at the right time and in the right place. A business that uses different digital marketing channels can interact with the target audience much more efficiently and in a timely manner, thereby constantly increasing the number of new customers and brand loyalty.

Marketing familiar to everyone has always involved huge resources of time and effort. It was difficult and of course, very expensive to track the process of implementation, promotion, and sale of the product. By contrast, digital marketing is simple, affordable, and incredibly effective. You can use several channels at once to build communication links with potential customers. You can communicate with them online, which allows you not only to quickly dispel all kinds of doubts about the products, but also to promptly answer questions of interest, building the necessary trust.

Now there is no need for mass mailings and tons of print runs, which, as a rule, go to waste, so digital marketing will significantly reduce the consumption of natural resources, while enhancing the efficiency of disseminating and promoting information. It is enough to create an advertising campaign, send it on the day of creation and immediately start tracking the effectiveness of its implementation. The necessary changes are made immediately, making it even better and more efficient.

**Literature review**

Digital marketing is the most effective step to success in any business; it is a completely innovative approach to the client offering new tactics, strategies based on a deeper understanding of user behavior in the network and in the market.

Digital marketing is developing faster than ever before around the world. In terms of advertising volume, digital promotion costs are steadily increasing. For example, in the United States alone, digital ad spending in 2020 was estimated at around $ 113 billion by 2020, double of what it had been just five years ago.

In Kazakhstan, the digital market is developing in the same trend of steady growth. According to the research company TNS Gallup, at the end of 2017 the volume of advertising on the Internet amounted to 6 billion tenge, and in 2018 the volume of the digital advertising market reached 9 billion tenge. However, these figures reflect, as a rule, only quantitative changes. Accordingly, questions arise about qualitative changes and what digital trends are most in demand today [1].

According to the monitoring of the advertising activity of the TNS agency, in 2017 the total advertising market was at the level of 41 billion tenge, which is a 13% increase as compared to the previous period. In 2018, this amount was already equal to 46 billion tenge, and even more is projected for the future.

Even though in the share of the media themselves, television is the leader with coverage of 72% of the audience, Kazakhstanis watch TV less compared to the previous year (81%). Next comes the Internet with 67% coverage [1].

If we talk about the advertising market from the point of view of an advertiser, it is logical that the share of online advertising is growing the most. Major international players such as Google (YouTube owner), Facebook and Yandex have the main money here. If we talk about the distribution share, then about 75% and 25% go to Google and Yandex, respectively [2].

Most people in Kazakhstan watch TV, this is more than 49% of the time, but at the same time, about 69% of advertisers' budgets are spent on television. The Internet is a different story. The share

in the budget today is closer to 13%, and the share of consumption over time is more than 35%. At the same time, Kazakhstani viewers spend about 197 minutes a day on TV, on the Internet they spend about 153 minutes per user [3].

Another trend is the growing impact of influencers in the country. Now one of the most popular lines in the budget of advertisers is the cost of bloggers and influencers. Affiliate advertising among popular bloggers in Kazakhstan is becoming one of the sales tools. Today blogging is a popular and profitable area that allows you to monetize this type of creativity not only for small and medium-sized businesses, but also for large international brands.

So, the emergence of digital marketing, the adaptation of classical marketing tools to the realities of Kazakhstan is unavoidable [4].

Thus, the relevance of research into digitalization of mechanisms for adapting classical marketing techniques lies in the construction and implementation of business models that will determine specific digital ways and methods. Promotion should be more effective, require less investment and generate more income, which will significantly boost the companies' competitiveness.

**The scientific novelty of the research** lies in the formation of innovative strategies and the development of methods for the transition from the traditional marketing to digital through a comprehensive theoretical and methodological study of the traditional promotion tools and techniques and their digitalization.

Elements of scientific novelty:

✓ Indication of the major directions in the transformation of promotion strategies in the digital economy;

✓ Testing foreign methods in the field of transition from traditional marketing to digital and assessment of their impact;

✓ Exploration of the basic methods of functioning of traditional and digital marketing based on the most innovative approaches in the context of the digital economy;

✓ Developing recommendations to improve the application of digital marketing elements and strengthen the links between business and the environment.

**Research methods and ethical issues**

The main research methods will be theoretical and experimental methods based on the search and scientific justification of the results of statistical data collected by the project team or provided by the national statistical agencies. Comparative analysis will also be widely used in the study of successful strategies for the development and implementation of digital marketing in different countries and in the formation of recommendations based on foreign experience. Empirical methods of marketing research will be widely used, namely, conducting quantitative and qualitative research, as well as the desk and field research, including cabinet – collection and processing of existing data from various sources and resources; field – quantitative – various types of surveys, etc.; qualitative – observations, interviews, etc.; and statistical ones.

Considering the novelty of the scientific direction, the project group will unify the theoretical definitions of digital marketing methods and tools and their classification, in order to form a logical construction of a reasonable structure of a marketing strategy for the transition to a digital standard. Also, statistical methods, methods of modeling and data visualization will be widely used.

**Expected results**

The section displays the following information:

1) the implementation of publications in peer-reviewed scientific journals (whether the results of scientific research carried out within the framework of the project are expected to be published and in which journal):

a) 1 article published, accepted for publication or submitted to a peer-reviewed scientific publication included in the Social Science Citation Index or Arts and Humanities Citation Index of

the Web of Science database, or having a Cite Score percentile in the Scopus database of at least 25 (twenty-five);

b) 2 articles in domestic or foreign journals recommended by the Committee for Quality Assurance in Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan (KOKSON) for publication of the main results of scientific research.

2) there is no possibility of patenting the results obtained in foreign patent offices (European, American, Japanese);

3) there is no possibility of patenting the results obtained in the Kazakh or Eurasian patent office, of the conclusion of a license agreement about intellectual property.

4) the expected scientific and socio-economic effect: for the first time, a methodological guide will be developed for the transition from traditional marketing to digital marketing.

5) the applicability of the obtained scientific results: the theoretical and practical results of the developed project will be used in government bodies, enterprises, universities of economic orientation, etc.

6) target consumers of the results obtained: domestic companies, state bodies of economic management, universities of economic direction, enterprises, and other scientific organizations.

7) influence on the development of science and technology – the results of the project make it possible to stimulate the development and growth of the country's economy.

**Conclusion**

An analysis of the current situation with a proposal for innovative promotion methods, strategies for business solutions will help bring the country's economy to a new level of digitalization of the economy.

The theoretical significance of the study lies in the fact that the results of the study on the development and transition to conceptually and technologically new methods of digital promotion will ultimately contribute to the development of the country's economy. The practical significance of the study lies in the development of a set of transition measures and new opportunities and alternatives to existing conventional methods of promotion, which will help attract the interest of entrepreneurs who could bring economic benefits by successfully competing with traditional approaches in business.

REFERENCES

1. Digital Marketing: What Digital Marketing Is and Why Your Business Needs It Today. [Electronic resource] URL: https://lafounder.com/article/digital-marketing (accessed: 10/25/2020)
2. Baghdasaryan, R. What is digital marketing. [Electronic resource] URL: https://actualmarketing.ru/marketing/chto-takoe-czifrovoj-marketing/ (accessed 08/25/2019)
3. Shevchenko D.A. (2017). Marketing education in Russia. - M.: Unity-Dana. P. 221.
4. Kotler F. (2019). Marketing from A to Z. 80 Concepts Every Manager Should Know. - M.: Alpina Publisher, S. 35.
5. Meyerson M. (2013). Internet Marketing Basics: Everything You Need To Know To Open Your Online Store. - M.: Mann, Ivanov and Ferber, p. 54.
6. Feoktistova, O. Digital marketing - what is it? [Electronic resource] URL: https://blog.ringostat.com/en/digital-marketing-chto-eto/ (accessed 09/02/2020)

**Омарова А.Ш., Алимжанова Л.М., Таштамышева А.Э.**

**Дәстүрлі маркетингті цифрлық форматқа ауыстыру әдістерін зерттеу және әзірлеу**

**Аңдатпа.** Цифрландыруға көшу – бұл қазіргі әлемдегі жаңа және қажетті бағыт, оған мемлекеттер мен компаниялардың бәсекеге қабілеттілігі, сондай-ақ азаматтардың өмір сүру сапасы тәуелді. Цифрлық маркетинг кәсіпорындардың және жалпы экономиканың мүмкіндіктерін кеңейтеді, оның барлық жүйелері мен құраушыларын нығайтады. Мұның

бәрі бейімделген инновациялық технологиялары бар цифрлық кеңістіктің заманауи инфрақұрылымы ұсынған кең мүмкіндіктерді пайдалана отырып, көп арналы цифрлық маркетингке көшудің әдістемесін әзірлеу қажеттілігін растайды, оның ішінде тауарды немесе қызметті нарыққа жылжытудан басқа, мақсатты аудиторияға нақты назар аудару, фокус-топты зерттеу және тұтынушының әзірленген профилі негізінде ары қарай ретаргентингпен тұтынушылардың реакциясын талдау. Сондықтан инновациялық маркетингтік құралдарды кеңінен енгізе отырып, маркетингтік компаниялардағы стандартты тәсілдерден цифрлық форматқа көшудің теориялық зерттеулерін және практикалық әдістемесін әзірлеу жоспарлануда.

**Түйінді сөздер**: дәстүрлі маркетинг, цифрлық маркетинг, жылжытылу, таргеттеу, жайғасым, әлеуметтік желілер.

**Омарова, А.Ш., Алимжанова, Л.М., Таштамышева, А.Э.**

**Исследование и разработка методов перехода традиционного маркетинга в цифровой формат**

**Актуальность.** Переход к цифровизации — это новое и необходимое направление в современном мире, от которого зависит конкурентоспособность государств и компаний, а также качество жизни граждан. Цифровой маркетинг расширит возможности предприятий и экономики в целом, укрепит все ее системы и составляющие. Все это подтверждает необходимость разработки методики перехода к многоканальному цифровому маркетингу с использованием предоставляемых современной инфраструктурой цифрового пространства с адаптированными инновационными технологиями широких возможностей, включающих, помимо продвижения товара или услуги на рынок, четкой направленности на целевую аудиторию, исследование фокус-групп, анализ реакции потребителя с последующим ретаргетингом на основе разработанного профиля клиента. Поэтому предполагается проведение теоретических исследований и разработка практической методологии перехода от стандартных подходов в маркетинге к цифровому формату с широким внедрением инновационного маркетингового инструментария.

**Ключевые слова**: традиционный маркетинг, цифровой маркетинг, продвижение, таргетирование, позиционирование, социальные сети.

**Авторлар туралы мәлімет:**

**Омарова Айгуль Шамилевна**, Іскери басқару докторы, Халықаралық ақпараттық технологиялар университеті, «Экономика және бизнес» кафедрасының қауымдастырылған профессоры.

**Алимжанова Лаура Муратбековна**, т.ғ.к., Халықаралық ақпараттық технологиялар университеті, «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры.

**Таштамышева Аида Эдуардовна**, Халықаралық ақпараттық технологиялар университеті, «Экономика және бизнес» кафедрасының магистранты.

**Сведения об авторах:**

**Омарова Айгуль Шамилевна**, доктор Делового Администрирования, ассоциированный профессор кафедры «Экономика и Бизнес», Международный университет информационных технологий.

**Алимжанова Лаура Муратбековна,** к.т.н., ассоциированный профессор кафедры «Информационные системы», Международный университет информационных технологий.

**Таштамышева Аида Эдуардовна**, магистрант кафедры «Экономика и Бизнес», Международный университет информационных технологий.

**About the authors:**

**Aigul Sh. Omarova**, Doctor of Business Administration, Associate Professor, Department of Economics and Business, International Information Technology University.

**Laura M. Alimzhanova**, Cand. of Tech. Sci., Associate Professor, Department of Information Systems, International Information Technology University.

**Aida E. Tashtamysheva**, master student, Department of Economics and Business, International Information Technology University.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ


| | |
|---|---|
| Ответственный за выпуск | Есбергенов Досым Бектенович |
| Редакторы | Далабаева Айсара Касымбековна<br>Джоламанова Балия Джалгасбаевна<br>Медведев Евгений Юрьевич |
| Компьютерная верстка | Туратауова Айжаркын Ахметовна |
| Компьютерный дизайн | Туратауова Айжаркын Ахметовна |




Подписано в печать 26.06.2021 г.
Тираж 500 экз. Формат 60x84 1/16. Бумага тип.
Уч.-изд.л. 10.1. Заказ №165

Издание Международный университет информационных технологий
Издательский центр КБТУ, Алматы, ул. Толе би, 59