Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Национальный исследовательский университет ИТМО»

(Университет ИТМО)

Факультет прикладной информатики

Образовательная программа Мобильные и сетевые технологии

Направление подготовки 09.03.03 Мобильные и сетевые технологии

О Т Ч Е Т

об учебной, ознакомительной практике

Тема задания: «Анализ и визуализация сведений о публикациях в журналах»

Обучающийся: Сапожников А.А., студент гр. K3240

Руководитель практики от университета: Валитова Ю.О., к.п.н., доцент ФПИн

Санкт-Петербург,

2025

СОДЕРЖАНИЕ

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

УДК (=UDC) – Universal Decimal Classification – общепризнанная систематизация научной литературы

ВАК – высшая аттестационная комиссия

ISSN – International Standard Serial Number – уникальный международный номер, идентифицирующий периодическое печатное или цифровое издание.

ВУЗ – высшее учебное заведение

DAX – Data Analysis Expressions – язык запросов для Power BI

РИНЦ – российский индекс научного цитирования

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Excel - программа для работы с электронными таблицами

Power BI - комплексное программное обеспечение для бизнес-анализа

Power Query - инструмент для извлечения, трансформации и загрузки данных в Power BI

PRIMORY_KEY – уникальный первичный ключ в реляционных базах

Индекс Хирша – это наукометрический показатель значимости научных исследований.

Дашборд – информационная панель, которая получает данные из других систем и отображает их в понятном виде.

Эффект Рингельмана – снижение продуктивности и эффективности работы группы при увеличении числа её участников.

ВВЕДЕНИЕ

Современная научная деятельность неразрывно связана с публикационной активностью, которая служит важным показателем продуктивности исследователей и качества научных изданий. В условиях постоянного увеличения объема научных данных актуальной задачей становится их систематизация и визуализация, позволяющие выявлять ключевые тенденции, анализировать динамику и принимать обоснованные решения. Данная учебная практика была посвящена сбору и анализу данных о журналах, входящих в перечень Высшей аттестационной комиссии (ВАК), их авторах, а также созданию интерактивных дашбордов для наглядного представления полученной информации.

Цель практики — разработка инструмента визуализации данных о журналах ВАК и их авторах с использованием Power BI, что способствует упрощению анализа публикационной активности в научной среде.

Основные задачи работы:

1. Сбор и структурирование информации о журналах ВАК (название, тематика, год основания) и их авторах (география, количество публикаций).
2. Очистка данных от дубликатов, устранение пропусков и приведение информации к единому формату
3. Создание интерактивных визуализаций в Power BI для анализа ключевых метрик: распределения авторов по регионам, динамики публикаций, рейтинга журналов по активности
4. Интерпретация полученных данных для выявления закономерностей в научной коммуникации

Актуальность работы обусловлена необходимостью автоматизированного анализа крупных массивов научной информации. Визуализация данных в Power BI позволяет ускорить обработку данных и сделать результаты анализа доступными для различных групп пользователей — от исследователей до администраторов научных организаций.

Методы исследования:

- Сбор данных из открытых источников (сайты журналов, базы данных eLibrary).
- Обработка информации с помощью Excel и Power Query.
- Создание дашбордов в Power BI с использованием карт, гистограмм и динамических фильтров.

Структура отчета включает теоретическое обоснование применяемых методов, описание этапов работы, анализ полученных результатов, а также рекомендации по дальнейшему совершенствованию процесса анализа научных данных. Данный отчет демонстрирует, как современные инструменты визуализации могут быть использованы для повышения эффективности работы с научной информацией, что особенно важно в условиях цифровизации науки и образования.

1. Участие во встречах с представителями профильных организаций и студентами старших курсов

Во время прохождения учебной практики мне довелось поучаствовать в двух встречах с представителями работодателей — Ильёй Николаевичем Горбуновым, генеральным директором IT-компании «Дата Аквизишн», и Валерией Евгеньевной Артамоновой, старшим системным аналитиком компании Ecom.tech.

На первой встрече Илья Николаевич представил успешные проекты своей компании, связанные с разработкой интеллектуальных систем, включая решения на основе технологий компьютерного зрения и обработки естественного языка. Встреча позволила глубже понять практическое применение технологий искусственного интеллекта, сложности, возникающие при их внедрении, и необходимые профессиональные навыки специалистов в этой области.

Вторая встреча была посвящена роли системного аналитика. Валерия Евгеньевна поделилась своим опытом работы в сфере аналитики, обсудила ключевые компетенции, важные для успешной карьеры, особенности взаимодействия с командой проекта и значение системного анализа в процессе разработки программного обеспечения.

Оба мероприятия предоставили уникальную возможность узнать больше о реальной работе в IT-сфере, а также помогли лучше осознать требования к профессионалам в областях разработки ИИ-решений и системного анализа.

2. Сбор и подготовка данных

2.1. Поиск журналов для сравнения

Сбор и структуризация данных всех журналы каждой категории требует больших временных затрат, поэтому, в первую очередь, перед командой стояла задача, отобрать по 2 журнала из каждой категории, которые будут максимально релевантны для анализа.

Было принято решение, оценивать сопоставимость журналов по УДК: действительно, журналы с одинаковым УДК имеют одну и ту же профессиональную специфику, тогда можно с высокой долей уверенности утверждать, что у этих журналах одинаковая целевая аудитория. Однако по каким-то причинам журнал А находится в категории К2, а журнал В в категории К3. Сравнив оба журнала, можно выявить объективные факторы, которые мешают журналу В быть в категории К2

В нашем распоряжении был перечень рецензируемых изданий, в котором были указаны УДК каждого журнала. Мы отобрали только те издания, УДК которых включает в себя УДК нашего журнала (2.3.4 и 5.2.3). Количество журналов, которые соответствуют требованиям из каждой категории представлено в Таблице 1

Категория	Количество точных совпадений УДК	Количество примерных совпадений УДК
К1	3	7
К2	2	12
К3	1	7

Таблица 1 – Количество журналов по УДК

В категории К3 возникла проблема: количество точных совпадений было меньше 2. Для разрешения вопроса мы решили добавить к списку журналы с “примерным” совпадением по УДК. Напомню, что наш журнал имеет УДК 2.3.4 и 5.2.3, тогда “примерно” похожие УДК имеют 2.3.X и 5.2.Y, где X и Y – некие положительные числа. Этот принцип основан на специфике систематизации УДК, в которой каждое новое число после каждй новой точки

уточняет предметную область. Количество “примерных” совпадений указано в Таблице 1. Таким образом, к журналам К3 был добавлен “Вестник современных цифровых технологий”.

2.2. Источники информации и обработка null-значений

После того, как стало ясно над какими журналами предстоит работать, пришло время определиться с источниками информации. В этом деле нам сильно помог сайт elibrary.ru - крупнейший российский информационно-аналитический портал в области науки, технологии, медицины и образования, содержащий рефераты и полные тексты более 38 млн научных публикаций. 90% информации мы взяли именно с этого ресурса. Также для расшифровки УДК мы использовали сайт perviy-vestnik.ru.

В этом разделе хотелось бы поговорить про ситуации, когда информацию найти не удавалось. Иногда в научных статьях использовались странные и громоздкие УДК, которые не поддавались дешифровке. Связано это, в первую очередь, с тем, что УДК ставили сами авторы и, в большинстве случаев, он не проверялся на соответствие с общепризнанной базой. В таких случаях мы просто сокращали номер УДК с конца до тех пор, пока дешифровщик не показывал результат. Для дальнейшего анализа сверхточное определение УДК было бы излишним

Сложнее, дело обстояло с авторами, у которых не было профиля на elibrary.ru, соответственно, чтобы узнать нужную информацию пришлось бы копаться на сайтах ВУЗов и научных издательств, что в разы бы уменьшило скорость обработки данных. Прежде чем решать возникшую проблему, было бы полезно оценить её масштаб: в процессе сбора информации авторы без профиля на elibrary пропускались. После того, как мы получили итоговые таблицы, выяснилось, что подавляющее большинство авторов без профиля писали статьи в соавторстве с более опытными коллегами. Выборочный поиск подтвердил то, что люди без профиля были в основном магистрантами или аспирантами, то есть не имели большого опыта в публикации научных статей.

Тогда, было принято решение не рассматривать этих авторов в дальнейшем анализе, вместо этого использовать данные их опытных коллег-соавторов

Но была ещё одна проблема - некоторые статьи были написаны исключительно авторами без профиля. К счастью, их было не много – около 3% от всех статей. Возможно, этими данными было разумнее пренебречь, но мной был предложен другой подход: добавить собирательный образ начинающих авторов с минимальными параметрами и сделать его автором всех статей, авторы которых не имели профиля на сайте. Безусловно, этот метод неидеален (например, получается, что этот собирательный образ написал десятки статей: этот результат нельзя учитывать в анализе наравне с обычными людьми), но он позволяет использовать все имеющиеся данные.

2.3. Описание базы данных

База данных состоит из 3 таблиц: “Авторы”, “Статьи” и “Связка”. Таблицы “Авторы” и “Статьи” имеют связь “многое ко многим” (то есть как статья может быть написана несколькими авторами, так и автор может написать несколько статей). В Power BI структура базы данных имеет следующую структуру: Рисунок 1.

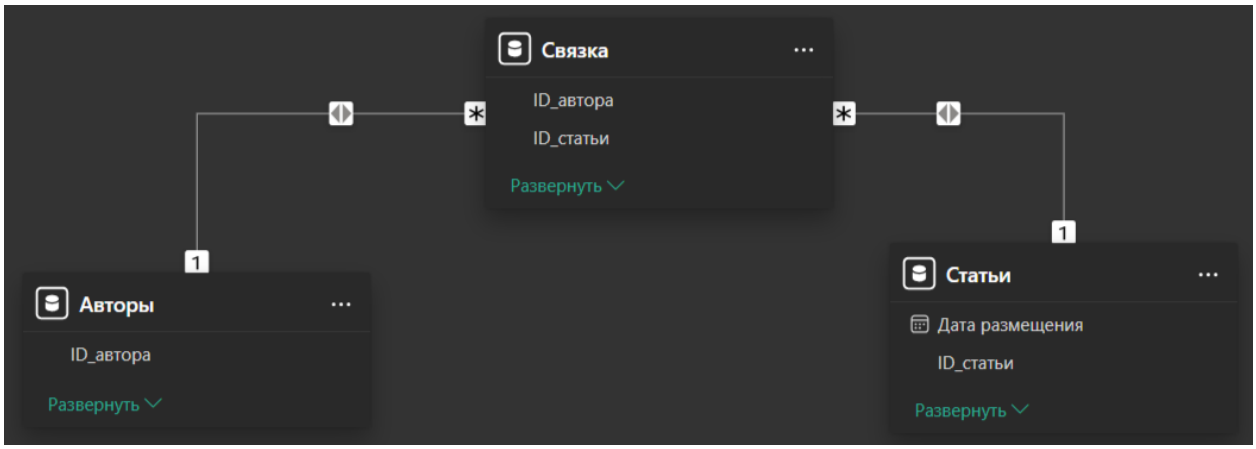

Рисунок 1 – структура базы данных

Таблица “Статьи” имеет следующие атрибуты

- PRIMORY_KEY: **ID_статьи** – [INTEGER] – уникальный ID из 11 цифр: первые 8 цифр – это ISSN, последние 3 цифры – порядковый номер
- **Категория** – [CHAR] – категория из перечня рецензируемых научных изданий, выбирается из списка (К1, К2, К3)

- **ISSN** – [CHAR] – уникальный номер журнала, выбирается из списка
- **Название статьи** – [VARCHAR] – полное название статей
- **Авторы** – [VARCHAR] – список авторов (этот скорее рудиментальный атрибут, который был создан до создания связи “многое ко многим”, этот атрибут не учавствует в анализе и существует чисто в ознакомительных целях)
- **УДК** – [INTEGER] – уникальный номер для каждой темы
- **Тема** – [VARCHAR] – дешифровка УДК
- **Ключевые слова** – [TEXT] – набор наиболее важных и часто втречающихся слов
- **Количество просмотров** – [INTEGER]
- **Количество уникальных загрузок** – [INTEGER]
- **Количество страниц** – [INTEGER]
- **Количество цитирований в РИНЦ** – [INTEGER]
- **Число персональных подборок пользователей** – [INTEGER]
- **Дата размещения** – [DATE]

Таблица “Авторы” имеют следующие атрибуты

- PRIMORY_KEY: **ID_Автора –** [INTEGER] – 4-значный уникальный номер, где первая цифра – номер категории, а последние 3 цифры – порядковый номер
- **ФИО** – [VARCHAR] – фамилия, имя, отчество
- **ВУЗ** – [VARCHAR] – полное официальное название ВУЗа
- **Число публикаций** – [INTEGER]
- **Число цитирований** – [INTEGER]
- **Индекс Хирша** – [INTEGER]
- **Год первой публикации** – [DATE]

Таблица “Связка” ставит в соответствие ID_статьи и ID_автора.

2.4. Инструменты обработки данных

Нам предстояло обработать большой массив данных: делать всё вручную было бы очень нерационально. С другой стороны писать большие скрипты, делать ML-модели, которые бы собрали информацию за нас потребовало бы от нас соответствующих компетенций и больших временных затрат. Поэтому мы комбинировали ручной поиск с скриптовой обработкой

Например, в самом начале пути перед нами встала нетривиальная проблема. Файл с перечнем рецензируемых научных изданий имел стобец “Научные специальности”, где были собраны вместе сплошным текстом номера УДК и название специальности. Более того, некоторые ячейки таблицы были разделены, чтобы информировать читателя о том, когда определённый УДК был присвоен журналу. Это ещё не всё: таблица занимала сотни страниц, и когда страница заканчивалась, таблица просто разбивалась в месте разрыва без какого-либо логического маркера. Простое преобразование в excel-формат ломало всю структуру файла, дальнейшее агрегирование не представлялось возможным.

Тогда я написал небольшой скрипт в Jupyter Notebook для обработки данных из pdf-файла, используя библиотеки (pdfplumber – для работы с pdf файлами, pandas – для преобразования в DaraFrame и re – для работы с регулярными выражениями).

```
import pdfplumber
import pandas as pd
import re

def match_both_specialties(text): # Функция для фильтрации (ищем и 2.3.X, и 5.2.Y)
    if text:
        has_23X = re.search(r'\b2\.3\.[1-8]\b', text)
        has_52Y = re.search(r'\b5\.2\.[1-8]\b', text)
        return has_23X and has_52Y  # Оставляем только если найдены оба шаблона
    return False

pdf_path = "Копия Перечень _на 09 12 2024.pdf"  # Открываем PDF
output_excel = "filtered.xlsx"
filtered_rows = []
current_row = None  # Переменная для объединения строк

with pdfplumber.open(pdf_path) as pdf:
```

```
    for page in pdf.pages:
        table = page.extract_table()
        if table:
            for row in table:
                if not row or len(row) < 3:  # Пропускаем пустые или некорректные строки
                    continue
                # Если строка начинается с числа (№ журнала) - это новая запись
                if row[0] and isinstance(row[0], str) and re.match(r'^\d+', row[0]):
                    if current_row:  # Если уже есть собранная строка, добавляем в список
                        filtered_rows.append(current_row)
                    current_row = row + [""] * (5 - len(row))  # Заполняем недостающие столбцы
                else:
                    if current_row: # Продолжаем объединять данные в текущую строку
                        for i in range(min(len(current_row), len(row))):  # Только пересекающиеся индексы
                            if row[i]:  # Добавляем только непустые значения
                                current_row[i] = (str(current_row[i]) + " " + row[i]).strip()
if current_row: # Добавляем последнюю обработанную строку
    filtered_rows.append(current_row)
# Фильтруем только журналы, где есть и 2.3.X, и 5.2.Y
filtered_rows = [row for row in filtered_rows if match_both_specialties(row[3])]
# Преобразуем в DataFrame
columns = ["№", "Наименование издания", "ISSN", "Научные специальности", "Дата включения"]
df = pd.DataFrame(filtered_rows, columns=columns) # Сохраняем результат в Excel
df.to_excel(output_excel, index=False, engine="openpyxl")
```

На выходе получаем удобный небольшой файл со всеми нужными журналами, данные по которым будут впоследствии визуализированы.

3 Анализ и визуализация данных

Данные собраны – пришло время их визуализировать. Но перед этим стоит сформулировать ряд гипотез, которые впоследствие с помощью визуализации в Power BI можно будет или доказать гипотезу или опровергнуть. Следующие разделы будут иметь похожую структуру: вместо заголовка будет название гипотезы, затем пояснение, визуализация и вывод.

3.1 Клиповое мышление

Утверждается, что статьи с меньшим количеством страниц пользуется бОльшим спросом. Оценивать “популярность” статей будем по совокупности критериев: количество пользовательских подборок, количество просмотров и количество цитирований. Визуализировать данные будем на совмещённом графике. По оси OY будет значения наших показателей, а по оси OX будет количество страниц разбитых по категориям (нам нет нужды привязываться к конкретным значениям страниц, поэтому будет логично разбить их на ранги с одинаковым шагом) соответсвующих таблице 2.

Категория	Количество страниц	Описание
A	<= 8	Короткая статья
B	9 – 12	Ниже среднего
C	13 – 16	Средний объём
D	17 – 20	Выше среднего
E	21 – 24	Большая статья
F	25 – 33	Огромная статья
G	>34	Аномально огромная статья

Таблица 2 – разделение статей по объёму

Чтобы програмно реализовать разделение на категории по объёму, написан следующий DAX-скрипт при создании столбца:

```
КатегорииОбъёма =
SWITCH(
    TRUE(),
    [Количество страниц] <= 8, "A<=8",
    [Количество страниц] <= 12, "B9-12",
    [Количество страниц] <= 16, "C13-16",
```

```
    [Количество страниц] <= 20, "D17-20",
    [Количество страниц] <= 24, "E21-24",
    [Количество страниц] <= 34, "F25-33",
    "G>34"
)
```

В процессе визуализации получились следующие графики (рисунок 2):

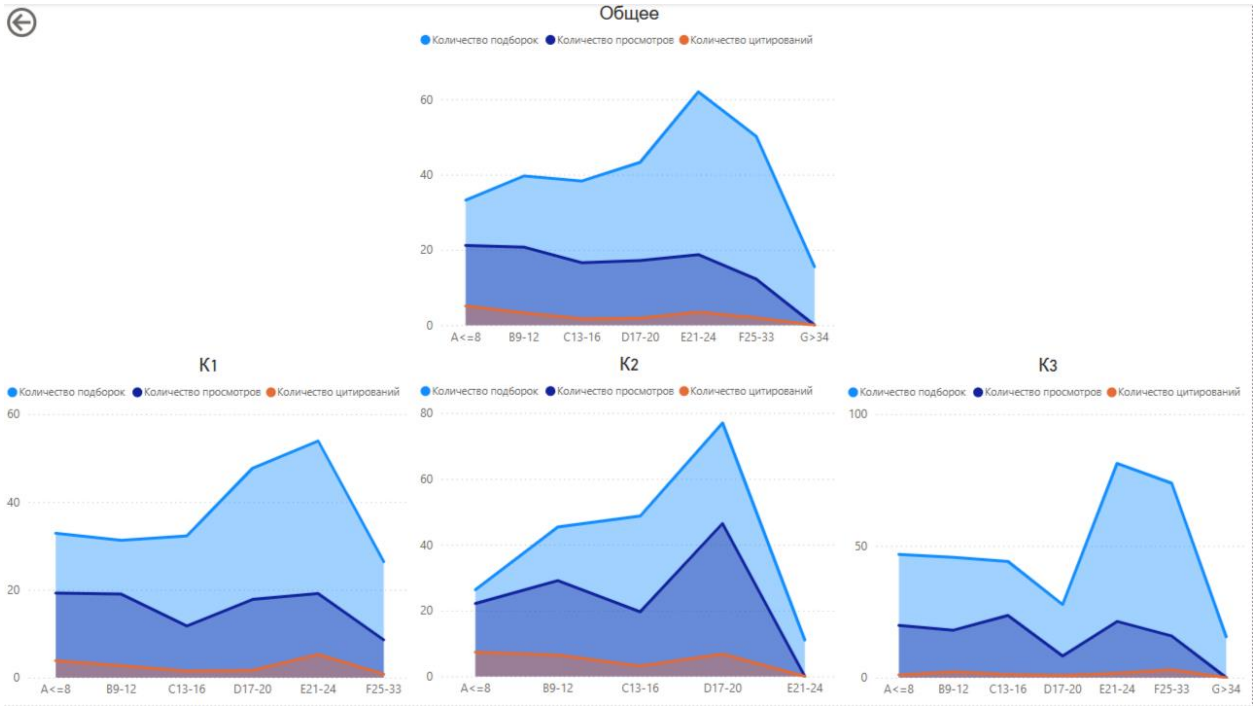

Рисунок 2 – Популярность в зависимости от объёма

Графики получились разными, однако можно выделить некоторые общие патерны:

- Глобальный максимум графиков находится в категории D-E (17-24 страницы) – это опровергает гипотезу о “клиповом мышлении”, иначе бы максимум достигался бы при меньших объёмах
- График имеет локальный максимум в категории B-C, между локальным и глобальным максимум явно можно видеть существенное снижение просмотров. Это наблюдение позволяет нам разделить читателей журналов на 2 группы: любителей, которые увлекаются наукой, но не готовы читать большие статьи, и дотошных экспертов, которых объём статей не пугает. Как мы выяснили в первом пункте: представителей второй группы больше
- Количество просмотров и персональных подборок неплохо коррелируют между собой. Это видно по поведению графиков. В

будущем под популярностью статей мы будем иметь один из этих параметров. Подробнее об этом в следующей главе.

- К2 сильно отличается от остальных категорий. В среднем журналы К2 имеют меньшее количество страниц, а потому в них несколько сдвинуты патерны, которые мы обсуждали выше. Скорее всего, это обусловлено нашим выбором, если бы вместо 2 журналов мы бы взяли все, то получили бы более усреднённую и предсказуемую картину.
- Статьи с экстремально большим объёмом не пользуются популярностью. Это важное замечание, так как все статьи >34 страниц взяты из журналов К3. Поэтому такие статьи нужно или сокращать, или разбивать на части. Читать 54 страницы одной статьи (реальный случай) будут только самые отчаянные.

Таким образом, можно сформулировать рекомендации по объёму статей:

1. Треть статей должны иметь объём от 9 до 12 страниц, остальные – от 17 до 24 страниц.
2. Статьи объемом больше 34 страниц нельзя допускать к печати

Безусловно, важнее не количество страниц, а их содержание, про это мы поговорим далее. Но формат изложения тоже важен: неоправданно большой или, наоборот, смехотворно маленький объём статьи может отпугнуть потенциальных читателей. Поэтому я посчитал нужным обратить внимание на этот аспект

3.2. Оценка экспертов и читателей

Утверждается, что оценка экспертов и оценка читателей коррелирует между собой. Сначала, нужно понять, какие характеристики можно принять за оценку. Нетрудно заметить, что если статья была процитирована в другой научной работе, то автор посчитал эту статью достаточно компетентной, чтобы на неё сослаться. Таким образом, можно использовать количество цитирований, как оценку от научного сообщества (экспертов). За оценку читателей возьмём количество пользовательских подборок, так как, вероятно, люди добавляют статью в свою подборку, потому

что посчитали эту статью небезынтересной. В качестве дополнительного параметра возьмём количество просмотров, он будет отвечать за диаметр кружков в точечной диаграмме. Получаем следующие графики: (рисунок 3)

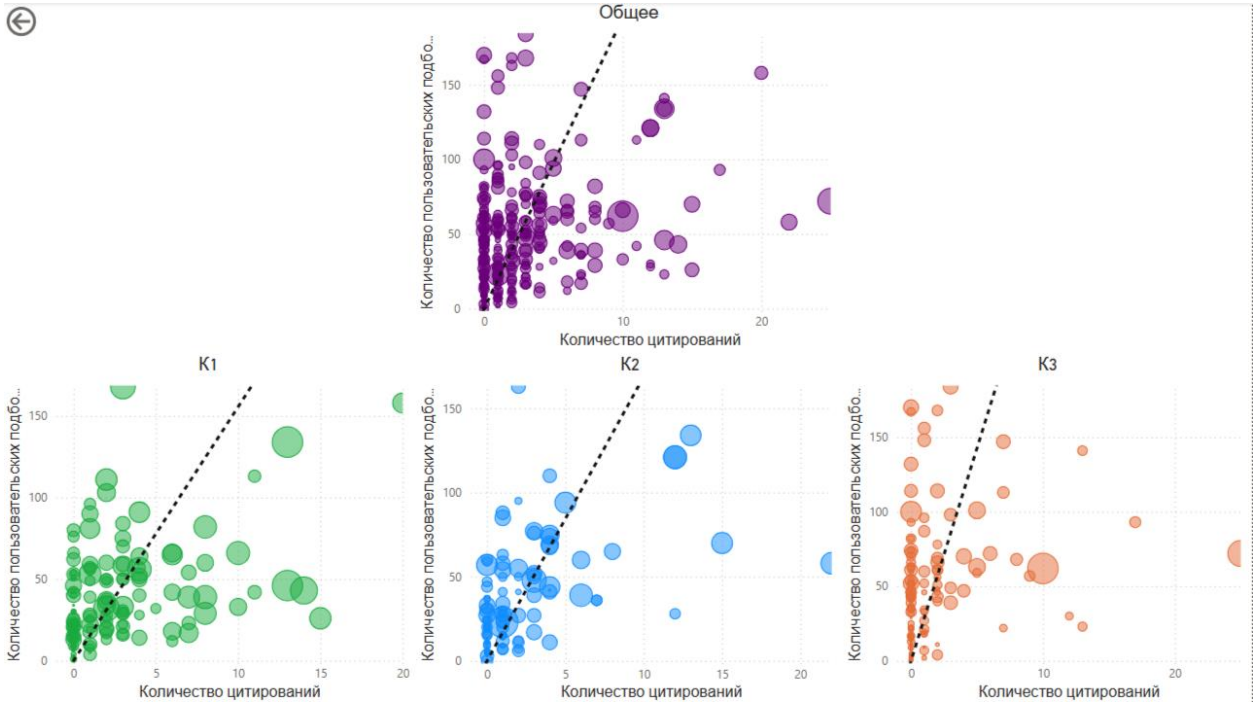

Рисунок 3 – Зависимость цитирований от подборок пользователей

По получившимся графикам нельзя говорить о явной корреляции между числом подборок и количством цитирований. В К1 корреляция почти отсутствует: есть много статей как с перевесом в пользу читателей, так и в пользу авторов, в К2 корреляция прослеживается более явно, а К3 далека об однозначности. Однако, это не значит, что из этих графиков нельзя сделать полезные выводы: если начертить линию соотношения, то у К3 угол наклона будет существенно больше, чем у К1 и К2. Это означает, что статьи К3 недооценены со стороны экспертов: существенная часть статей вообще не имеет цитирований, несмотря на популярность у пользователей. Из всего вышесказанного, можно дать следующую рекомендацию: *нужно работать над увеличением числа цитирований в РИНЦ*. Скромное количество цитирований сильно бросается в глаза при сравнении журналов, предположу, что это один из основных факторов, из-за которого научные журналы остаются в К3.

3.3. Опыт и индекс Хирша

Утверждается, что индекс Хирша коррелирует с опытом автора. Индекс Хирша – это (грубо говоря) показатель продуктивности учёного. Он зависит от количества научных статей и от количества цитирований этих статей. Однако, индекс Хирша не учитывает опыт автора в написании научных статей, хотя опыт научного сотрудника является одним из основных характеристик. В дальнейшем мы бы хотели получить некий параметр “качества” автора: если индекс Хирша коррелирует с опытом, то по нему можно судить о “качестве” автора, если индекс Хирша не будет коррелировать с опытом, придётся придумывать альтернативные оценки качества.

Для доказательства корреляции воспользуемся уже знакомым точечным графиком: по оси Ox будет год начала научной карьеры, по оси OY индекс Хирша. Получим следующие графики (рисунок 4):

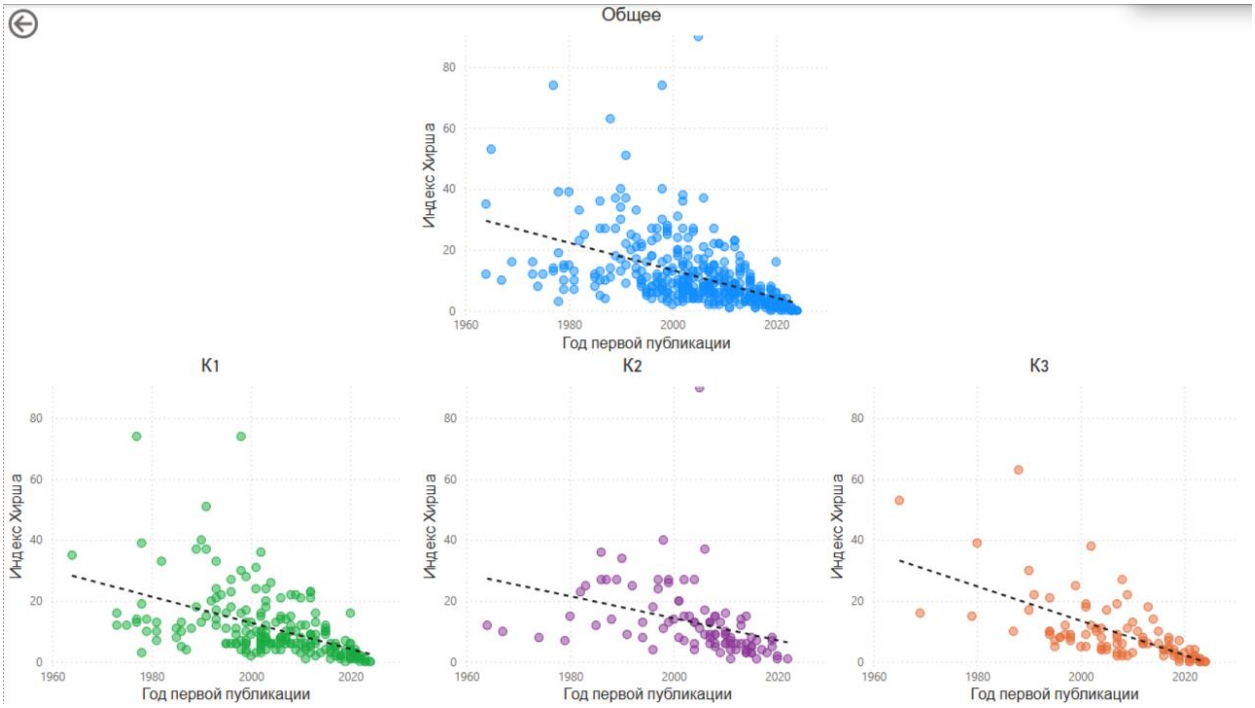

Рисунок 4 – Зависимость индекса Хирша от года первой публикации

На графике отчётливо видно, что опыт и индекс Хирша сильно коррелируют друг с другом. Да, есть выбросы, среди более опытных людей большой разброс, но основная масса авторов сосредоточена очень близко к линии тренда – это и подтверждает сильную связь между опытом и индексом Хирша.

3.4. Динамика изменений содержания выпусков

Хочется в разрезе посмотреть на содержание журнала и проследить динамику изменения “качества” статей. В предыдущей главе мы доказали, что параметр “качества” автора может быть измерен индексом Хирша. Показателем качества статьи будем считать суммарный индекс всех авторов статьи. Да, такой подход имеет ряд недостатков: например, именитые учёные могут плохо сработаться и суммарный вклад в общее дело будет существенно меньше, чем арифметическая сумма. С другой стороны, может произойти эффект синергии, когда команда специалистов в результате совместной работе покажет результат, который превосходит сумму вклада каждого поодиночке. Поэтому суммарный индекс Хирша будет неким усреднённым значением “качества” статьи.

Для каждой категории сделаем по 2 графика: первый будет накопительным и показывать доли в процентах, а второй будет отображать абсолютные числа. Самое время поговрить про категории, на которые мы разобъём все статьи за год (аналогично, как мы это делали в разделе 3.1). Они будут напрямую связаны с суммарным индексом Хирша (таблица 3)

описание я сам придумал для удобства, оно может не соответствовать действительности

Категория	Суммарный индекс Хирша	Описание
A	0 – 1	Начинающий автор
B	2 – 6	С небольшим опытом
C	7 – 12	Опытный учёный
D	13 – 20	Уважаемый учёный
E	21 – 35	Именитый учёный
F	>35	Ведущий учёный

Таблица 3 – разделение статей суммарного индексу Хирша

Теперь мы можем построить наши графики (рисунок 5)

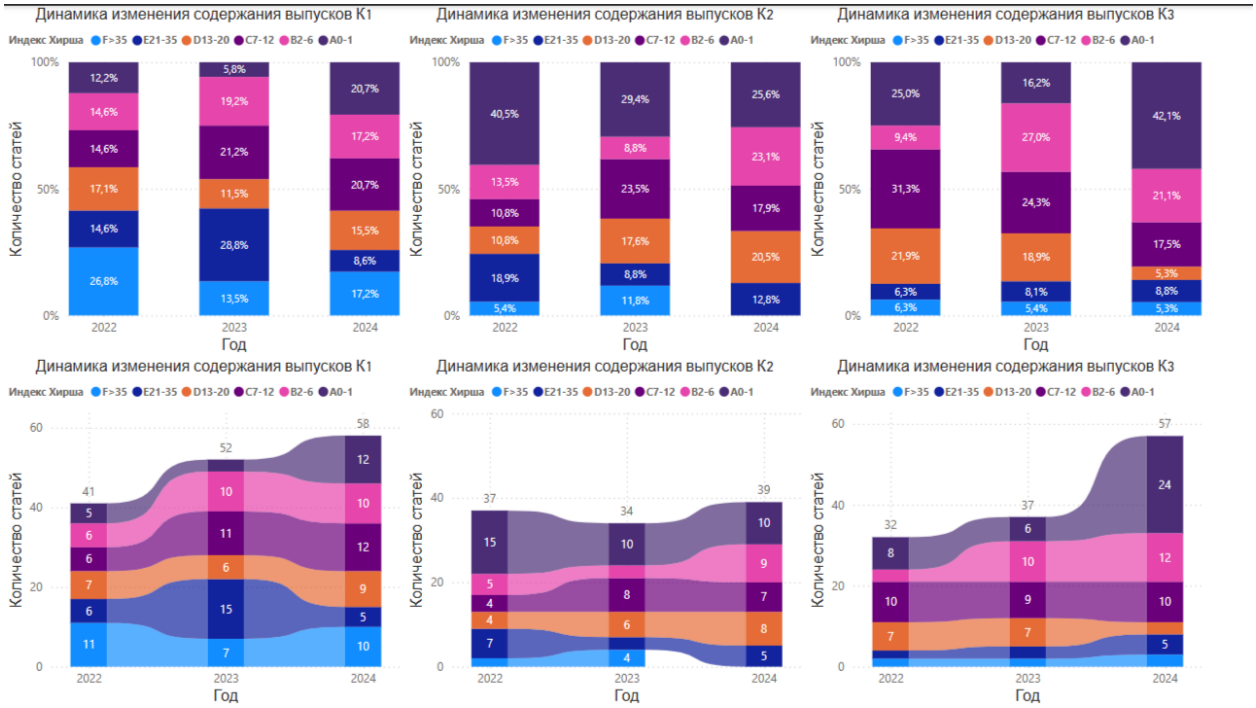

Рисунок 5 – Динамика изменения содержания выпусков

Получаем подробную и наглядную инфографику. Видно, что в К3 идёт резкий рост публикаций, но происходит этот рост преимущественно за счёт начинающих авторов. Количество статей с суммарным индексом Хирша меньше 2 достигает 42,1% - это вдвое больше, чем в других категориях. Обратите внимание на К1, количество публикаций растёт год от года, но мы можем наблюдать прирост и за счёт более опытных авторов, а не только за счёт дебютантов (хотя и их количество тоже растёт) – это вполне позитивная динамика. Логично сделать вывод, что для улучшения качества журнала нужно ужесточить отбор для статей.

Косвенным подтверждением может служить ситуация с журналами из К2, несложно заметить, что они стагнируют: количество публикаций практически не растёт, статьи с суммарным индексом больше 35 вообще не публикуются, но при этом журнал остаётся в К2. Возможно, причина в том, что количество статей начинающих авторов остаётся в пределах 25%.

Я ни в коем случае не хочу сказать, что статьи неопытных авторов не нужны, напротив, это дальновидный подход, когда университеты-учередители взращивают специалистов для своих нужд. Но и читатели хотят читать качественные статьи от именитых учёных. Именно поэтому большой процент

публикаций от неопытных авторов подрывает качество издания. Я могу предложить два выхода из сложившейся ситуации:

1. Ужесточить отбор, чтобы повысить качество статей. Самый простой путь.
2. Приобщать начинающих авторов работать вместе с более опытными колеегами. Это увеличит суммарный индекс Хирша и позитивно скажется на качестве публикаций.

Если идти по второму пути, будет полезно рассмотреть, как количество авторов влияет на качество статей.

3.5. Оптимальное количество авторов

Итак, перед нами стоит задача: выяснить сколько авторов нужно, чтобы написать наиболее успешную статью. Ответ нетривиален: малое количество людей могут не справиться с большим объёмом информации, большая группа людей может работать непродуктивно из-за эффекта Рингельмана.

Для начала нужно создать отдельный столбец в таблице “Статьи”, в котором будет содержаться количество авторов. Сделать это можно написав следующий скрипт в DAX:

```
Количество_авторов =
COALESCE(
CALCULATE(
    COUNT('Связка'[ID_автора]),
    FILTER(
        'Связка',
        'Связка'[ID_статьи] = 'Статьи'[ID_статьи]
    )
), 0)
```

За меру качества статьи, не измения традициям из прошлых разделов, возьмём количество цитирований. Получим следующий график для всех трёх категорий. (Рисунок 6).

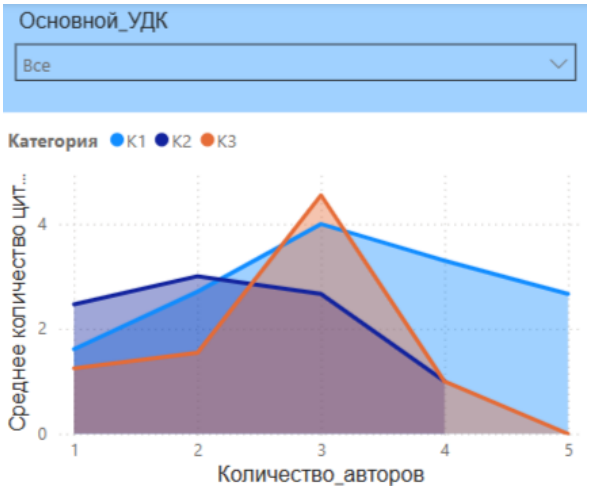

Рисунок 6 – Зависимость количества авторов от цитирований статьи

По графику чётко видно, что оптимальное количество авторов для написания статьи равно трём. Однако этот ответ может быть уточнён для каждой специализации отдельно, для этого вверху графика был установлен фильтр по УДК. С помощью него можно посмотреть оптимальное количество авторов конкретно для вашей специальности. Например, (рисунок 7)

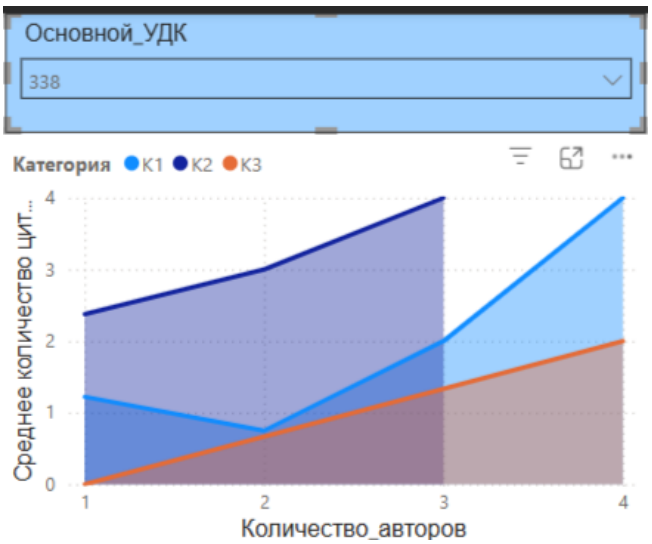

Рисунок 7 – пример для УДК 338

Выбрав номер УДК 338 (Экономическое положение. Экономическая политика) можно увидеть, что наилучший результат имеют статьи, написанные четырьмя авторами, а не тремя.

Однако УДК 338 является исключением из правила, для большинства специализаций именно при команде из трёх человек достигается максимальный результат. Это замечание стоит учитывать при формировании научно-иследовательских групп.

4. Дополнительные визуализации

В прошлом модуле мы разбирали те визуализации, по которым можно сделать некие глобальные выводы. Но в процессе работы были созданы ряд дашбордов, которые напрямую не дают ответа на наши вопросы. Однако не стоит умолять их важность: они раскрывают более мелкие и локальные аспекты анализа. Кроме того, они удобно визуализируют данные при помощи динамической составляющей BI систем.

4.1. Визуализация в рамках категории

Именно с этой визуализации начинался наш проект. 4 из 5 разделов из предыдущего модуля опираются на эту визуализацию. Она созданы по данным в рамках категории и даёт представление о локальных процессах. Для каждой категории своя визуализация, но структура везде одинаковая. Например, вот визуализация категории К1 (Рисунок 7)

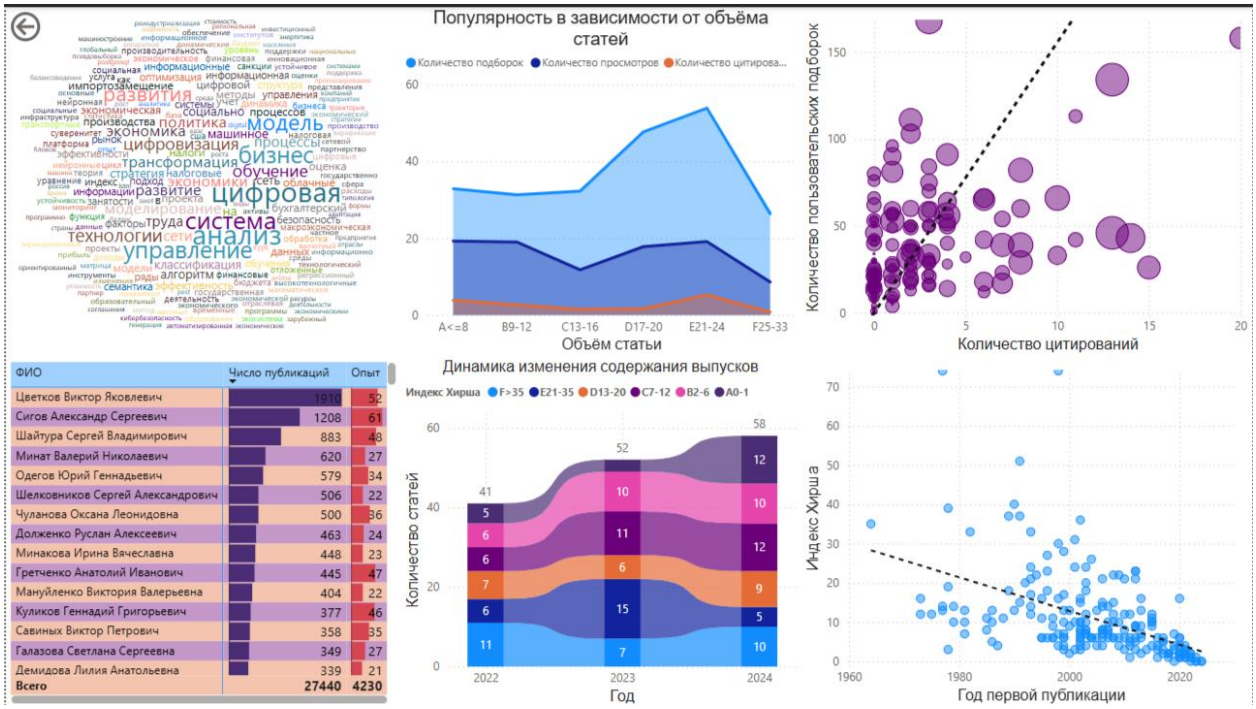

Рисунок 7 – визуализация категории К1

В центральнем и правом ряду находятся 4 визуализации из предыдущего модуля (3.1 сверху по центру, 3.2 сверху справа, 3.3 снизу справа, 3.4 снизу по центру). Вправо внизу находится совмещённая матрица с гистограммой, где содержится информация об авторах статей. Используя динамическую составляющую Power BI можно детально посмотреть на

локальные данные. Например, можно узнать какие люди имеют суммарный индекс Хирша от 21 до 35 в категории К2 и посмотреть их результаты (рисунок 8)

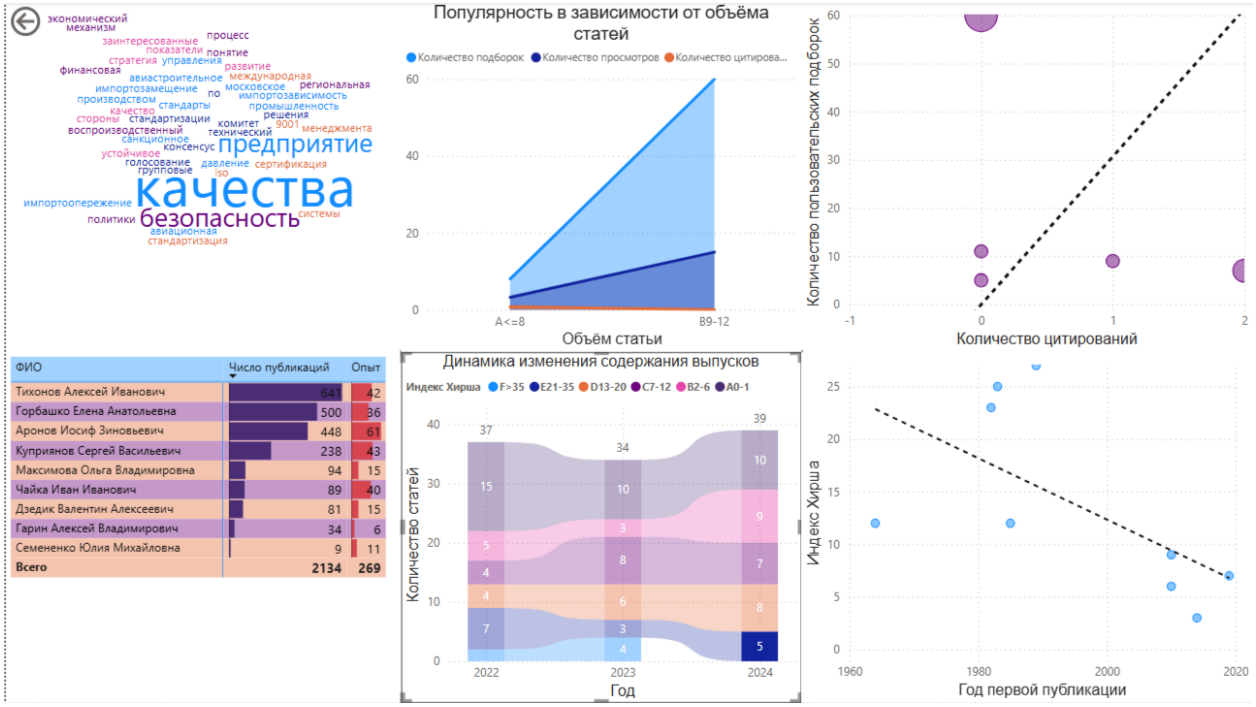

Рисунок 8 – информация по автором из категории Е

Или например, узнать результаты деятельности конкретного человека (Щербакова А.Ю.) из К3 (рисунок 9)

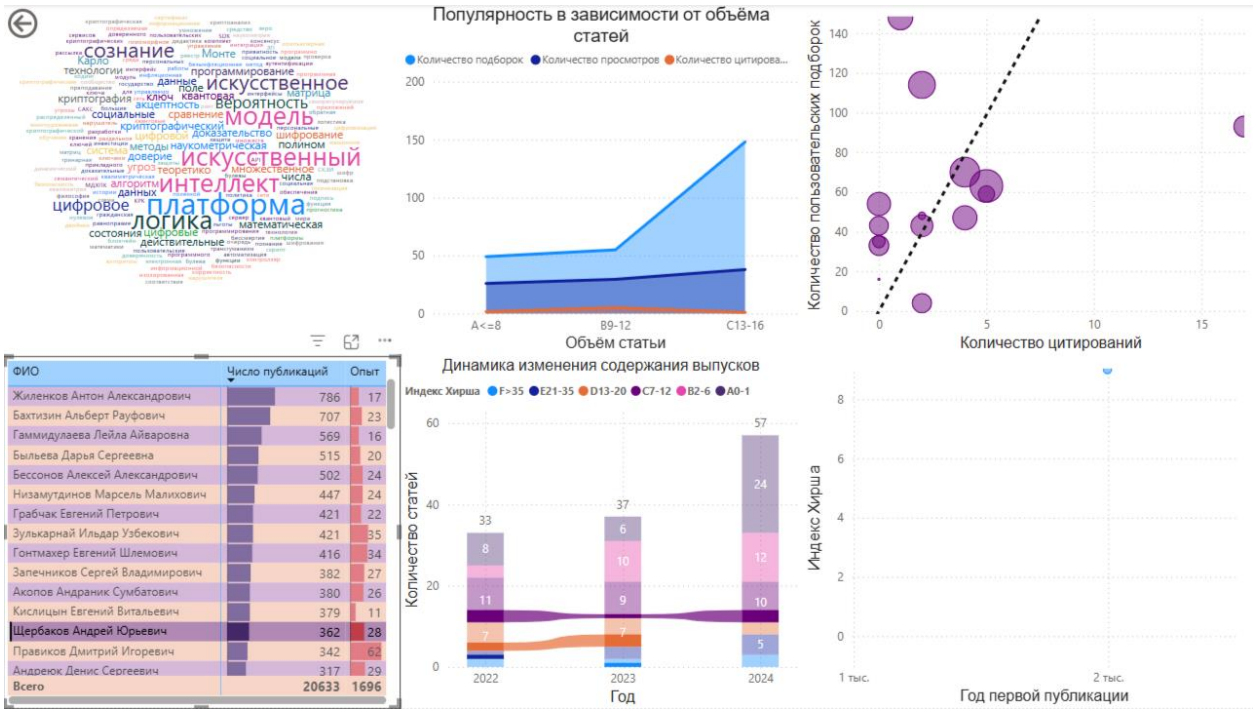

Рисунок 9 – информация о Щербакове А.Ю.

Глобальных выводов по этим дашбордам сделать трудно, но зато можно посмотреть данные более детально и составить гипотезы, которые будут проверяться на других дашбордах, которые были представлены в прошлом модуле.

4.2. Облака слов

В прошлом разделе я не рассказал про необычный график сверху слева. Этот график визуализирует вероятность встретить определённое слово в статьях данного раздела (чем чаще встречается слово, тем оно больше) – такой график называется облоком слов. Конечно, он обрабатывает не все слова из статьи, а только те, которые заявлены, как ключевые – они хранятся в таблице “статьи”. Этот график выходит за рамки стандартного инструментария Power BI, поэтому был скачан с внешнего источника.

Аналогично другим графикам, облако слов может помочь ответить на вопросы локального характера используя динамичность BI системы. Например, для того, чтобы отобрать информацию о статьях из К1, где используется слово “анализ” (рисунок 10)

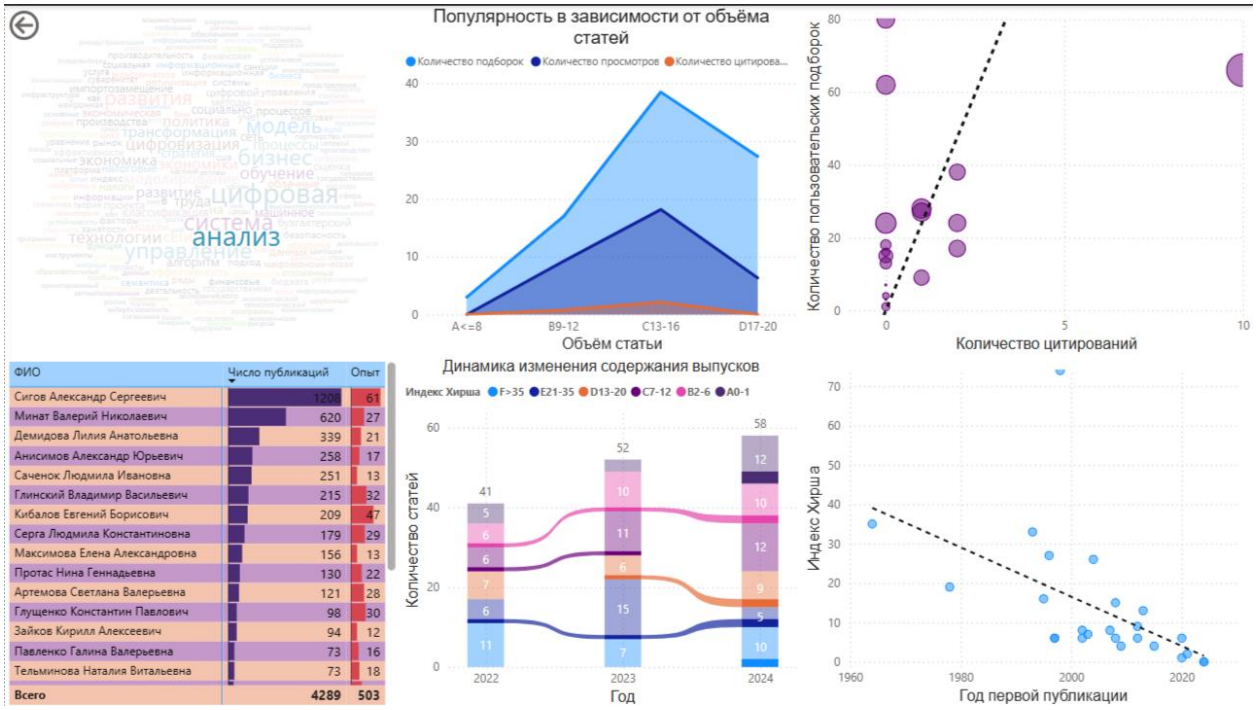

Рисунок 10 – информация по статьям, где фигурирует слово “анализ”

Также можно сделать дополнительный дашборд где можно сравнить облака слов по каждой категории (рисунок 11)

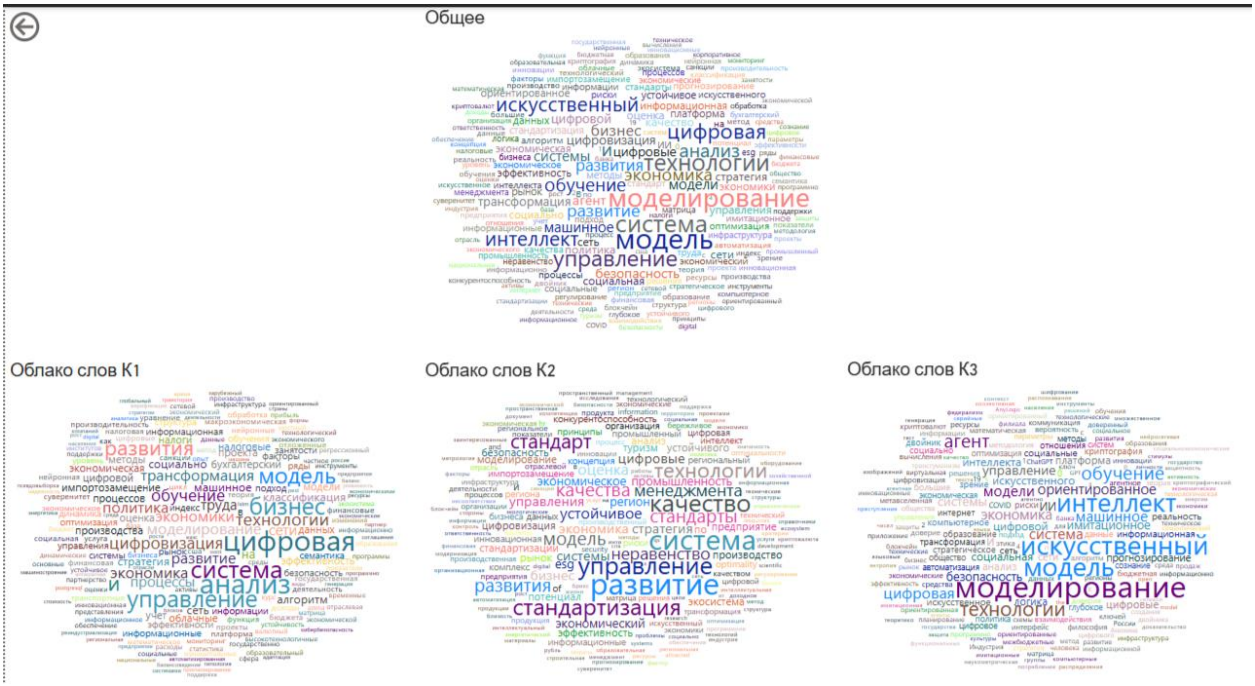

Рисунок 11 – Облака слов

Если в главном облаке слов выделить конкретное слово (например, “система”), то на других графиках будут отображаться слова, которые используются с этим словом в каждой категории соответственно (Рисунок 12)

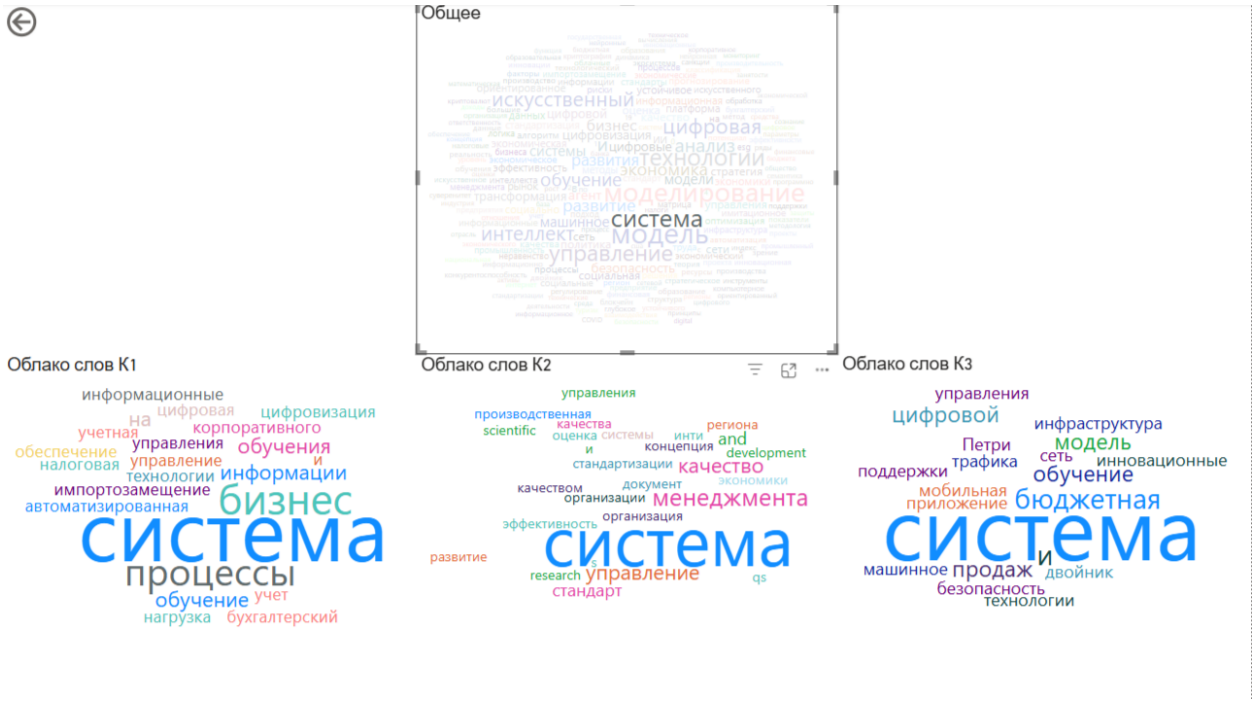

Рисунок 12 – облака слов относительно конкретного слова

Как уже было сказано ранее, такие визуализации нельзя назвать неинформативными: они могут раскрыть некоторые мелкие детали в данных, но вытащить из них что-то, что можно экстрапалировать на все данные будет затруднительно. Поэтому они и находятся ближе к концу отчёта.

ЗАКЛЮЧЕНИЕ

В рамках учебной практики была выполнена работа по сбору, структуризации и анализа данных с последующей визуализацией и разработке конкретных рекомендаций по улучшению научного журнала. В заключение, хочется собрать все рекомендации на одной странице

1. **Стандартизация по объёму**. Статьи должны быть примерно разделены на две категории: объёмом от 9 до 12 страниц и объёмом – от 17 до 24 страниц. Статьи объёмом больше 34 страниц крайне нежелательны к публикации
2. **Повышение цитирований в РИНЦ.** Нужно работать над этим показателем: писать качественные статьи на актуальные темы, которые захотят процитировать.
3. **Уменьшить количество статей от неопытных специалистов.** Доля статей с суммарным индексом Хирша меньше 2 должна быть не больше 25% (сейчас 40%). Молодых специалистов следует приобщить к делу по средствам менторства более опытных коллег
4. **3 – оптимальное количество соавторов.** Понятно, что в зависимости от специализации это число может меняться, но сильно отходить от этого показателя не стоит
5. **Увеличить количество освещаемых подтем в рамках выбранной специализации.** Этот пункт вытекает из отчётов моих коллег. Суть в том, что публикация статей под разнообразными УДК (в рамках определённого форватора) положительно влияет на популярность у слушателей

Стоит оговориться, что все эти выводы могут быть уточнены и скорректированы, если данные будут дополнятся.

В рамках учебной практике я приобрёл ценный опыт работы с BI-системами, сборе и структуризации данных, формировании и доказательств гипотез. Также нельзя забыть про мой опыт взаимодействия с моими коллегами, это существенно улучшает мои Soft Skills. Я рад, что смог

поработать над данным проектом. Уверен, что опыт, который я приобрёл в рамках учебной практике я смогу успешно применить в реальных аналитических проектах на моей будущей стажировке.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Каталог журналов elibrary (https://elibrary.ru)
2. Дешифратор УДК (https://perviy-vestnik.ru/udc/)
3. Портал ISSN (https://portal.issn.org/resource/ISSN/)
4. Показатели цитирования от Google (https://scholar.google.com)
5. Информация о кодах специализации (https://www.garant.ru)
6. Поиск информации о преподавателях (https://studizba.com/hs/rtu/teachers/)
7. Гайд по pandas и jupyter (https://tproger.ru/articles/gajd-po-obrabotke-dannyh-s-pomoshhyu-pandas-chast-pervaya)
8. Цикл роликов про Power BI от Your Mentor (*видеохостинг не доступен на территории РФ*)
9. Курсы от Microsoft по Power BI (https://learn.microsoft.com)
10. Гайд по DAX от Microsoft (https://support.microsoft.com/ru-ru/office)
11. ГОСТ для научно-иследовательских статей (https://science.itmo.ru/)