МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра Математического и компьютерного моделирования

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ИЗ SCOPUS И ОБРАБОТКА БИБЛИОГРАФИЧЕСКОЙ ИНФОРМАЦИИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТА

студента 5 курса 561 группы

направление 09.03.03 — Прикладная информатика

механико-математического факультета

Машков Никита Владимирович

Научный руководитель
зав. каф., д.ф.-м.н., доцент Ю.А. Блинков

Зав. кафедрой
зав. каф., д.ф.-м.н., доцент Ю.А. Блинков

Саратов 2020

Введение. В настоящее время существует проблема сбора библиографической информации для предоставления различных отчетов. В бакалаврской работе рассмотрен импорт библиографической информации из БД Scopus. При этом в качестве формата выходных данных был выбран широко известный JavaScript Object Notation (JSON), удобный для чтения человеком и компьютером. JSON-файл, содержащий библиографическую информацию, легко импортируется в документо-ориентированную БД MongoDB. MongoDB предназначена для гибкой, масштабируемой и очень быстрой работы даже при больших объёмах данных. Для работы с MongoDB был использован модуль pymongo языка программирования Python, позволяющий не только загружать информацию в базу данных, но и писать к ней различные запросы.

MongoDB представляет собой набирающую в последнее время популярность технологии NoSQL. При использовании NoSQL не требуется создавать несколько связанных друг с другом таблиц. Данные о таких разных публикациях, как статьи, книги, учебные пособия, тезисы и материалы конференций хранятся в одной коллекции (collection, аналог таблицы в реляционных базах данных (БД)). При этом у записей, соответствующих публикациям разных типов, в общем случае не совпадают поля. Например, для статьи нужно хранить наименование журнала, в котором она была опубликована, а для учебного пособия требуется хранить наименование издательства и т.д. При этом некоторые поля у записей совпадают (авторы, год издания и т. д.). При использовании реляционных БД пришлось бы создавать отдельную таблицу для публикаций каждого вида. В то же время применение технологии NoSQL позволяет отказаться от жёсткой структуры БД.

Актуальность исследования объясняется практической потребностью хранения библиографической информации о публикациях сотрудников СГУ.

Целью работы являются некоторые способы хранения библиографической информации и способы её обработки для дальнейшего использования, статистического анализа данных о публикациях сотрудников СГУ.

Структура бакалаврской работы. Бакалаврская работа состоит из введения, трех глав и заключения. В первой главе проведено сравнение различных систем управления библиографической информацией. Во второй главе описаны реляционные и нереляционные БД, проведено сравнение реля-

ционного и NoSQL подходов к хранению данных. В третьей главе описаны способы импорта библиографической информации с БД Scopus.

Основное содержание работы.

Организация работы с библиографической информацией является неотъемлемой частью научно-исследовательской работы преподавателей и студентов. Традиционные способы, такие как создание картотек уходят в прошлое. Развитие информационных технологий привело к созданию специализированного программного обеспечения – систем управления библиографической информацией (СУБИ). Архитектура современных СУБИ включает следующие компоненты:

1. базу данных для хранения библиографической информации (автор, название, издательство, журнал, год и т. д.);
2. интерфейс с библиотечными каталогами, онлайн-журналами и другими базами данных;
3. интерфейс с текстовыми процессорами, который позволяет автоматически вставлять библиографические ссылки и списки;
4. систему генерации ссылок и списков, которая позволяет сразу оформлять их в требуемом стандарте.

С конца 80-х годов двадцатого века фактическим стандартом баз данных (БД) стали реляционные базы данных (РБД) и системы управления реляционными базами данных (РСУБД).

В 90-х годах двадцатого века бурное развитие глобальных сетей и распределенных систем привело к потребности эффективно обрабатывать большие объемы данных. При этом данные были распределены по многим узлам сети. Для решения этой проблемы было предложено объединить компьютеры в кластеры, на которых хранились фрагментированные данные. К сожалению, эффективно решить поставленную задачу с помощью привычных РСУБД не удалось.

NoSQL – это не замена традиционного реляционного подхода к хранению и обработке данных. Данный подход применяется тогда, когда решаемые задачи связаны или с большими и постоянно возрастающими объемами данных (которые требуют высокой масштабируемости), или связанных с хранением таких данных, которые сильно отличаются от ре-

ляционной формы представления (документно-ориентированные системы, объектно-ориентированные системы). Часто создают СУБД с поддержкой как традиционно-реляционного подхода, так и альтернативных NoSQL-решений. Реляционные модели лучше подходят для относительно небольших объемов данных высокой ценности (таких как данные о пользователях некоторой информационной системы, билинговая информация), а NoSQL решения – для больших объемов данных низкой ценности (ведение логов и сбор статистики, хранение документов).

Scopus – это самая крупная в мире единая реферативная база данных (БД) о научных публикациях. В этой БД проиндексированы более 21000 наименований научно-технических и медицинских журналов, издаваемых примерно 5000 международных издательств. БД обновляется ежедневно. В ней содержатся все номера всех томов журналов, выпущенных ведущими научными издательствами. В БД содержатся ссылки на все вышедшие рефераты опубликованных научных статей.

БД Scopus наполнена научной литературой самого высокого качества. Некоторая часть публикаций находится в открытом доступе (Open Access). Помимо журнальных статей в Scopus включены также труды научных конференций и публикации, доступные только в электронной форме. Составной частью поисковой системы Scopus является Research Performance Measurement (RPM). RPM – это средства контроля эффективности исследований. Они помогают оценивать авторов, направления в исследованиях и журналы.

В настоящее время данные, содержащиеся в Scopus, используются Минобрнауки РФ в качестве критериев общероссийской системы оценки эффективности деятельности высших учебных заведений.

В БД Scopus содержатся:

1. 21,000 рецензируемых журналов (включая около 3,800 журналов Open Access и около 400 российских журналов).
2. 100,000 книг.
3. 390 наименований Trade Publications.
4. 370 книжных серий (продолжающихся изданий).
5. 6,8 млн. конференционных докладов из трудов конференций.

6. 50 млн. записей: 29 млн. записей со ссылками с 1996 г. (из которых 84% включают пристатейную литературу).
7. 21 млн. записей с 1996 г. и до 1823 г.
8. 27 млн. патентных записей от пяти патентных офисов Статьи в предпечатной подготовке («Articles-in-Press») доступны из более 3,850 журналов.

Согласно данным компании Elsevier преимущества Scopus перед другими БД заключаются в следующем:

1. Scopus превышает по полноте и ретроспективной глубине большинство существующих в мире баз данных;
2. в Scopus содержится полная информация по отечественным научным организациям, журналам и авторам (приведены показатели цитируемости и др.);
3. в Scopus есть специально разработанные средства для того, чтобы контролировать эффективность проводимых исследований, оценивать авторов, организации, направления в исследованиях и журналы;
4. нет эмбарго, многих рефераты индексируются в БД ещё до выхода печатного варианта;
5. Scopus обладает удобным и простым для конечного пользователя интерфейсом;
6. различные варианты написания названия журнала, фамилии и имени автора, названия организации можно при необходимости объединить.

Для того, чтобы эффективно анализировать публикации, содержащиеся в БД Scopus, компания Elsevier предоставляет заинтересованным разработчикам возможность работы с сайтом через его интерфейс прикладного программирования (API). Для работы с API Scopus программист должен получить на сайте ключ API (API key), по которому его можно будет впоследствии идентифицировать. Также организация, в которой трудится разработчик, должна оплатить подписку на БД Scopus.

Каждый ключ API имеет определенные ресурсы, квоты и уровни обслуживания, включенные по умолчанию. Ниже в соответствии с таблицей 1 приведены параметры API Scopus для тех, у кого есть подписка, и тех, у кого её нет.

Доступ к определенным интерфейсам не включен по умолчанию, потому что они требуют дополнительного разрешения от Elsevier.

В случае, когда вызов API абстрактного поиска происходит из сети, настроенной для доступа к Scopus, тогда ответ соответствует столбцу «Есть подписка». Когда тот же самый вызов поступает с анонимного IP-адреса, ответ соответствует столбцу «Нет подписки».

Квоты сбрасываются каждые 7 дней. Неиспользованная часть оставшейся квоты и дата сброса квоты доступны в HTTP-заголовках ответа API.

Таблица 1 — Параметры API Scopus

Название API	Есть/Нет	Нет подписки	Есть подписка	Недельная квота	Запрос/ сек
Serial Title	Есть	STANDARD, COVERIMAGE views / 25 результатов (макс. 200)	STANDARD, COVERIMAGE, ENHANCED 25 результатов (макс. 200)	20000	3
Citations Count Metadata	Нет	N/A	STANDARD view / 25 результатов (макс. 200)	50000	18
Citations Overview	Нет	N/A	STANDARD view / 25 результатов (макс. 200)	20000	3
Subject Classifications	Есть	Нет ограничений	Нет ограничений	N/A	N/A
Abstract Retrieval	Есть	META view	All views, default FULL view	10000	6
Affiliation Retrieval	Есть	N/A	All views, default STANDARD view	5000	6
Author Retrieval	Есть	N/A	All views / Max 25 results	5000	3
Affiliation Search	Есть	N/A	25 результатов (макс. 200)	5000	3
Author Search	Есть	N/A	25 результатов (макс. 200)	5000	3
Scopus Search	Есть	STANDARD view / Default 25 results	STANDARD view / Макс. 200 результатов COMPLETE view / Макс. 25 результатов COMPONENT view / Макс. 25 результатов	20000	6

Приведём основные особенности вышеуказанных API Scopus:

Serial Title API представляет собой интерфейс для поиска периодических изданий по названию или ISSN.

Abstract Citations Count API представляет собой интерфейсы для извлечения количества ссылок на документы SCOPUS. Данный API возвращает изображение с водяными знаками или метаданные в форматах JSON / XML. Метаданные также содержат дополнительные идентификаторы документов, если они доступны.

Citation Overview API представляет собой интерфейс для извлечения количества ссылок на документы SCOPUS с разбивкой по годам с возможностью исключения самоцитирования. Чтобы получить общее количество цитирований для документа, нужно использовать Abstract Citations Count API.

Subject Classifications API обеспечивает возможность поиска тематических классификаций, связанных с контентом ScienceDirect или Scopus.

Abstract Retrieval API представляет собой интерфейс для получения аннотации документа Scopus. Полный реферат содержит множество метаданных, в том числе ссылки на профили авторов и организаций, в которых они работают. Текст аннотации доступен для поиска с использованием Scopus Search API.

Результат запроса выдаётся в формате XML, и, хотя его части могут быть переведены в JSON, сложность разметки затрудняет перевод аннотации в формат JSON.

Content Affiliation Retrieval API представляет собой интерфейс для получения профиля организации в SCOPUS. Ответ может содержать ссылки на Scopus Search и Author Profiles. Профили организаций индексируются и могут быть найдены с использованием Affiliation Search API.

Author Retrieval API представляет интерфейс для получения профиля автора в Scopus. Документ может содержать ссылки на Scopus Search и Affiliation Profiles. Профили автора индексируются и могут быть найдены с использованием Author Search API.

Affiliation Search API представляет собой интерфейс поиска, связанный с профилями организаций. Каждый результат поиска по определению будет ссылаться на профиль организации.

Author Search API представляет собой интерфейс поиска профилей того или иного автора в Scopus. Каждый результат поиска по определению будет ссылаться на профиль автора. У записей поиска также могут быть ссылки на текущий профиль автора.

Scopus Search API представляет собой интерфейс поиска аннотаций статей в Scopus. В результатах поиска также могут быть ссылки на полный текст статьи.

Преимущества использования API Scopus заключаются в следующем:

1. Прямой доступ к данным Scopus в реальном времени.
2. Архитектура RESTful: независимая, масштабируемая, переносимая и надежная платформа.
3. Поддержка стандартов и спецификаций: W3C CORS, Dublin Core, PRISM.
4. Простота интеграции с клиентскими приложениями и / или напрямую с клиентскими веб-сайтами.
5. Разнообразие поддерживаемых форматов ответов API.
6. Ответы API включают ссылки на соответствующие ресурсы для упрощения навигации и доступа.
7. Документация интерактивного API позволяет просматривать запросы и ответы API в любом из поддерживаемых форматов ответов непосредственно с портала разработчика Elsevier.

Для импорта библиографической информации из Scopus была написана программа на языке Python. Для каждого автора из следующего списка персональных идентификаторов сотрудников кафедры математического и компьютерного моделирования:

```
MiKM = [
  ('6701893186',  'Блинков Юрий Анатольевич'),
  ('34880252800', 'Кондратов Дмитрий Вячеславович'),
  ('57205741334', 'Иванов Сергей Викторович'),
  ('56288336500', 'Крылова Екатерина Юрьевна'),
  ('7005930757',  'Кузнецова Ольга Святославовна'),
  ('56205546400', 'Панкратов Илья Алексеевич'),
]
```

производится GET-запрос к БД Scopus следующего вида:

```
api_resource = "https://api.elsevier.com/content/search/scopus"
params = {
  'field': 'dc:identifier,citedby-count',
}
headers = {
  'X-ELS-APIKey': my_api_key,
  'X-ELS-ResourceVersion': 'XOCS',
  'Accept': 'application/json',
}

count, cites, res = 0, 0, {}
for id, fio in MiKM:
  articles = []
  for y in range(1990, 2025, 5):
    params['query'] = "AU-ID(%s) AND PUBYEAR > %d AND (PUBYEAR < %d OR PUBYEAR = %d)"
    response = requests.get(api_resource, params=params, headers=headers)

    author = response.json()
    if response.status_code == 200 and 'error' not in author['search-results']['entry
      articles += author['search-results']['entry']
```

Здесь произведена разбивка по годам с интервалом 5 лет это связанно с тем, что для нулевого уровня доступа ограничено число возвращаемых записей в количестве 25. Это сильно усложняет логику запросов и может не решит проблему, когда в одном году опубликовано больше 25 статей данного автора.

Этот запрос возвращает JSON-объект, в котором содержатся номера статей, написанных данным автором. Затем для каждой статьи делается ещё один запрос к БД Scopus:

```
  api_resource = "https://api.elsevier.com/content/abstract/scopus_id/" + scopus_id
  params = {
    'field': 'authors,title,publicationName,volume,issueIdentifier,prism:pageRange,cov
  }
  headers = {
    'X-ELS-APIKey': my_api_key,
    'X-ELS-ResourceVersion': 'XOCS',
    'Accept': 'application/json',
  }

```

```
18    response = requests.get(api_resource, params=params, headers=headers)
19    assert response.status_code == 200
20    paper = response.json()['abstracts-retrieval-response']
```

На этот раз в возвращённом JSON-объекте содержится информация об именах всех авторов статьи (это работает только с компьютеров из сети университета, поскольку по их IP адресам определяется доступ в Scopus); журнале, в котором она опубликована; номере журнала; номерах страниц и типе публикации (статья в журнале, статья в трудах конференции и т.д.). Также доступна аннотация к статье и номер DOI (если он у неё есть).

Ниже приведен результат выполнения программы

```
      Блинков Юрий Анатольевич:   33 (299)
Кондратов Дмитрий Вячеславович:   18 (54)
      Иванов Сергей Викторович:    2 (0)
     Крылова Екатерина Юрьевна:   19 (50)
 Кузнецова Ольга Святославовна:    4 (6)
     Панкратов Илья Алексеевич:    2 (0)
                         Итого:   78 (409)
```

Урезанный результат экспорта информации из MongoDB:

1. Chelnokov Y.N., Pankratov I.A., Sapunkov Y.G. Optimal reorientation of spacecraft orbit. Archives of Control Sciences, 2014, vol. 24, no. 2, pp. 119-128.
2. Blinkov Y.A., Gerdt V.P., Pankratov I.A., Kotkova E.A. Construction of a New Implicit Difference Scheme for 2D Boussinesq Paradigm Equation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, vol. 11661 LNCS, pp. 152-163. DOI: 10.1007/978-3-030-26831-2_11
3. Kuznetsova O.S., Tkachev V.G. Asymptotic properties of solutions to the hele-shaw problem. Doklady Akademii Nauk, 1999, vol. 367, no. 2, pp. 164-165.
4. Kuznetsova O.S., Tkachev V.G. Asymptotic properties of solutions to the Hele-Shaw equation. Doklady Mathematics, 1999, vol. 60, no. 1, pp. 35-36.
5. Kuznetsova O.S., Tkachev V.G. Length functions of lemniscates. Manuscripta Mathematica, 2003, vol. 112, no. 4, pp. 519-538. DOI: 10.1007/s00229-003-0411-3

6. Kuznetsova O.S. Polynomial solutions to the Hele-Shaw problem. Siberian Mathematical Journal, 2001, vol. 42, no. 5, pp. 907-915. DOI: 10.1023/A:1011963510477
7. Yakovleva T.V., Bazhenov V.G., Krysko V.A., Krylova E.Y. Contact interaction plates, reinforced by ribs, with gaps under the influence of white noise. PNRPU Mechanics Bulletin, 2015, vol. 2015, no. 4, pp. 259-272. DOI: 10.15593/perm.mech/2015.4.15
8. Awrejcewicz J., Krylova E.Y., Papkova I.V., Krysko V.A. Regular and chaotic dynamics of flexible plates. Shock and Vibration, 2014, vol. 2014 DOI: 10.1155/2014/937967
9. ...

Заключение. Было разработано программное обеспечение, позволяющее импортировать библиографическую информацию из БД Scopus и экспортировать в различных форматах. В качестве примера БД MongoDB была заполнена публикациями сотрудников кафедры математического и компьютерного моделирования по состоянию на апрель 2020 г. Приведены примеры различных запросов для обработки накопленной информации.