Поиск патентов-аналогов на основе сравнения ключевых понятий

С.А. Фоменков, Д.М. Коробкин, В.С. Коробкина

Волгоградский государственный технический университет

Аннотация: В данной статье описаны подходы к автоматизации полнотекстового поиска по ключевым фразам в области патентной информации. Автоматизация поиска по ключевым фразам (n-граммам) является существенно более сложной задачей, чем поиск по отдельным словам, кроме того требует проведения морфологического и синтаксического анализа текста. Для достижения поставленной цели были решены следующие задачи: (а) проанализированы системы полнотекстового поиска: Apache Solr, ElasticSearch и ClickHouse; (б) проведено сравнение архитектур и основных возможностей каждой системы; (в) получены результаты поиска в Apache Solr, ElasticSearch и ClickHouse на одном и том же наборе данных. Были сделаны следующие выводы: (а) все рассмотренные системы осуществляют полнотекстовый поиск по ключевым фразам; (б) Apache Solr является системой с самой высокой производительностью, также у неё максимально удобные функции; (б) ElasticSearch обладает быстрой и мощной архитектурой; (в) ClickHouse имеет высокую скорость обработки данных.
Ключевые слова: поиск, ключевые фразы, патент, Apache Solr, Elasticsearch, ClickHouse.

Введение

Эксперты патентного ведомства осуществляют анализ текущего уровня техники и поиск аналогов регистрируемого изобретения в основном на базе той информации, которую выдают специализированные поисковые системы [1,2] в патентной области. При этом извлечение ключевых слов и фраз из текста заявки на изобретение эксперт производит вручную.

Целью данной работы является анализ систем полнотекстового поиска (по ключевым фразам). Автоматизация поиска по ключевым фразам (n-граммам) является существенно более сложной задачей, чем поиск по отдельным словам, кроме того требует проведение морфологического и синтаксического анализа текста. Для достижения поставленной цели необходимо решить следующие задачи:

- Изучить и проанализировать системы полнотекстового поиска: Solr, ElasticSearch и ClickHouse;
- Сравнить архитектуры и основные возможности каждой системы;

- Сравнить результаты поиска в Solr, ElasticSearch и ClickHouse на одном и том же наборе данных.

Анализ СУБД для полнотекстового поиска

Apache Solr [3] — это сверхбыстрая мультимодальная поисковая платформа с открытым исходным кодом, построенная на возможностях полнотекстового, векторного и геопространственного поиска Apache Lucene [4].

Архитектура:

– Ядро Solr

– Lucene: библиотека для полнотекстового поиска, на базе которой построен Solr.

– ZooKeeper [5]: используется для управления кластером Solr

Основные возможности и особенности:

– Репликация базы данных — процесс копирования данных из основной базы данных в одну или несколько баз данных-реплик для улучшения доступности и надежности данных.

– Сегментирование базы данных — стратегия горизонтального масштабирования, которая выделяет дополнительные узлы или компьютеры для распределения рабочей нагрузки приложения.

– Мощные функции поиска.

– Анализ текста.

– Разнообразные API для интеграции с другими системами и приложениями.

– Интерфейс администрирования.

ElasticSearch [6] — распределенная поисковая и аналитическая система RESTful [7] с открытым исходным кодом, масштабируемое хранилище данных и векторная база данных, способная удовлетворить растущее число вариантов использования. Являясь сердцем Elastic Stack [8], система

централизованно хранит данные для быстрого поиска, точной настройки релевантности и мощной аналитики, которая легко масштабируется.

Архитектура:

– Elasticsearch Node

– Lucene

– Kibana [9]

Основные возможности и особенности:

– Горизонтальное масштабирование.

– Репликация.

– Поиск в реальном времени.

– Анализ данных на высокой скорости и в масштабе для обеспечения наблюдаемости, безопасности и поиска с помощью Kibana. Мощный анализ любых данных из любого источника: от анализа угроз до поисковой аналитики, журналов, мониторинга приложений и многого другого.

– API.

ClickHouse [10] — это система управления базами данных с открытым исходным кодом, ориентированная на столбцы, которая позволяет создавать отчеты с аналитическими данными в режиме реального времени.

Архитектура:

– Колоночная СУБД.

– Масштабируемая архитектура.

– Параллельная обработка.

Основные возможности и особенности:

– Высокая производительность.

– SQL поддержка.

– Хранение данных.

– Масштабируемость.

– Интеграция.

Тестирование возможностей СУБД

Было выбрано для тестирования 5 патентов, в текстах 2 содержится ключевая фраза «медицинская техника» в различных падежах:

- Релаксирующее устройство на основе нагревательных элементов (yandex.ru/patents/doc/RU216836U1_20230302).
- Массажёр для стоп (yandex.ru/patents/doc/RU2033781C1_19950430) (Рис. 1).
- Корсет для вытяжения позвоночника (yandex.ru/patents/doc/RU189471U1_20190523).
- Многофункциональное женское платье (yandex.ru/patents/doc/RU134010U1_20131110).
- Корсет (yandex.ru/patents/doc/RU2520039C1_20140620).


```
},{
  "id":"2033781",
  "title":["МАССАЖЕР ДЛЯ СТОП"],
  "number":["2033781"],
  "publication_date":["1995-04-30T00:00:00Z"],
  "priority_date":["1991-06-06T00:00:00Z"],
  "authors":["Карминский В.Д. (RU)","Соломин В.А. (RU)","Филь Е.С. (RU)","Калинченко С.Ю. (RU)"],
  "patent_holders":["Карминский Валерий Давидович","Соломин Владимир Александрович","Филь Евгений Сергеевич","Калинченко
  "references":["SU1692577A1, A61H 23/00, 1988"],
  "abstract":["Использование: для массажного воздействия на конечности. Сущность изобретения: массажер содержит основание
  "claims":["МАССАЖЕР ДЛЯ СТОП, содержащий основание, выполненное в виде герметичной эластичной оболочки, заполненной жид
  "description":["Изобретение относится к медицинской технике и предназначено для использования в медицинских и спортивны
  "ipc_classes":["A61H 15/00(2006.01)"],
  "_version_":1804928399946285056
},{
```


Рис. 1. – Описание патента на примере «Массажёр для стоп»

Apache Solr можно загрузить с официального сайта по ссылке, обязательно требуется предустановленный Java Development Kit [11].

Необходимо проиндексировать патентные данные в формате xml, внесенные в базу данных Apache Solr. Мы будем индексировать патенты в формате xml.

Введём запрос «медицинская техника» для всех атрибутов. В итоге на тестовой выборке было найдено 2 патента (Рис. 2).

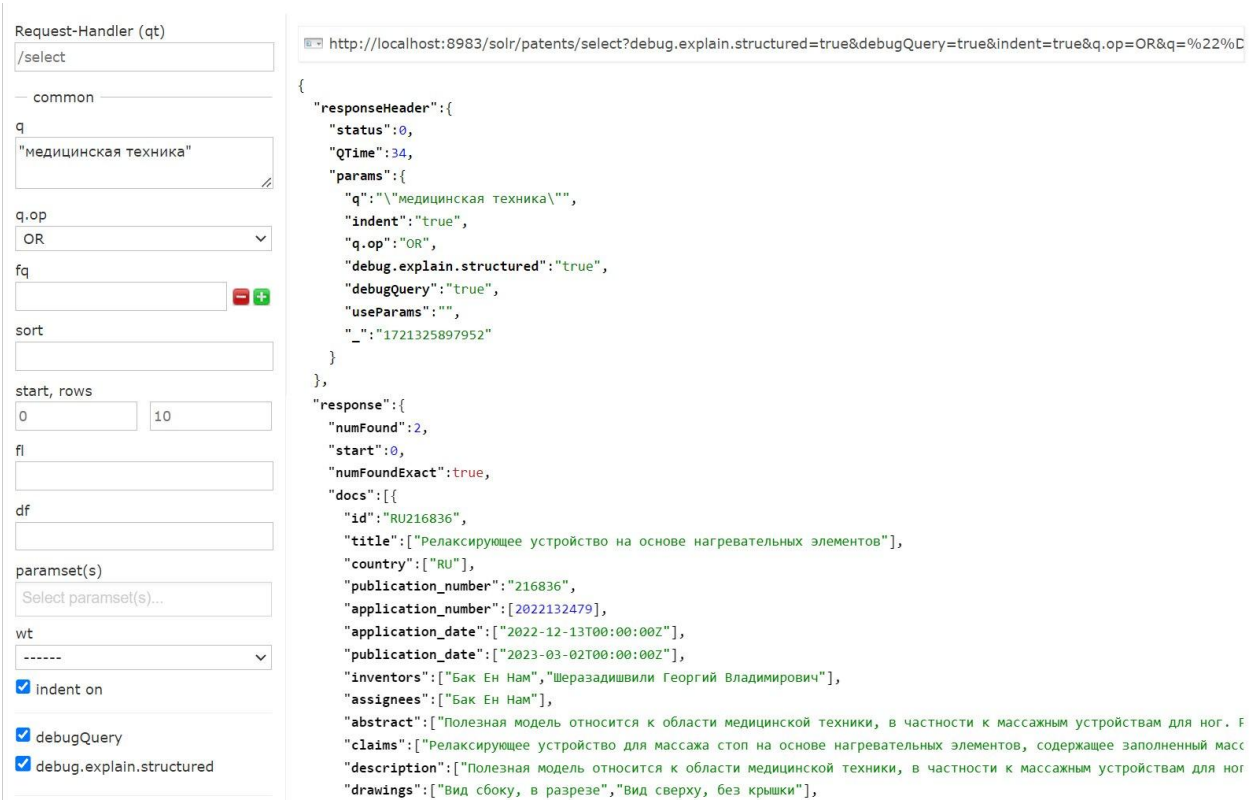

Рис. 2. – Поиск в системе Solr

Загружаем zip-файл ***ElasticSearch*** для Windows с официального сайта, а также zip-файл Kibana для Windows. Далее проиндексируем предоставленные патентные документы в ElasticSearch. Т.к. стандартный фильтр русского языка в Elasticsearch работает неидеально, скачиваем zip-файл с словарём hunspell [12], распаковываем его в папку config внутри ElasticSearch. С помощью команды PUT создаём индекс с патентами.

С помощью команды POST /patents/_doc заполним индекс патентами (Рис. 3).

Введём запрос «медицинская техника» для всех атрибутов. Выбираем запрос типа «multimatch», чтобы была возможность искать сразу в нескольких полях. В «query» пишем те слова, которые нужно найти, в «operator» – and (тогда будут показывать патенты, содержащие все слова, а не одно из нескольких). В «fields» пишем атрибуты патента, в которых ищем слова, в «source» атрибуты найденных патентов, которые будут отображаться.

```
# Заполним индекс патентами
POST /patents/_doc
{
    "id": "RU216836U1",
    "title": "Релаксирующее устройство на основе нагревательных элементов",
    "country": "RU",
    "publication_number": "216836",
    "application_number": "2022132479",
    "application_date": "2022.12.13",
    "publication_date": "2023.03.02",
    "inventors": [
        "Бак Ен Нам",
        "Шеразадишвили Георгий Владимирович"
    ],
    "assignees": "Бак Ен Нам",
    "abstract": "Полезная модель относится к области медицинской техники, в
      частности к массажным устройствам для ног. Релаксирующее устройство
      на основе нагревательных элементов содержит бокс, заполненный
      массирующим наполнителем в виде песка или песочной смеси,
      нагревательные элементы, устройство управления нагревательными
      элементами, панель управления нагревательными элементами, выполненная
      с возможностью выбора программы, времени и температуры работы
      нагревательных элементов. В корпусе бокса размещены нагревательные
      элементы, соединенные с устройством управления нагревательными
      элементами и панелью управления нагревательными элементами.
      Технический результат заключается в повышении эффективности
      воздействия на стопы ног в процессе проведения теплотерапии,
```

Рис. 3. – Заполнение индекса патентами

В итоге на тестовой выборке было найдено 3 патента (Рис. 4-5).

```
POST /patents/_search
{
  "query": {
      "multi_match" : {
        "query" : "медицинская техника",
        "operator": "and",
        "fields": ["id","title", "country","publication_number"
          ,"application_number","publication_number","publication_date"
          ,"application_date","inventors","assignees","abstract","claims"
          ,"description","drawings","similar_documents","related_documents"],
        "fuzziness": "AUTO"
      }
  },
  "_source": ["id","title", "country","publication_number","application_number"
    ,"publication_number","publication_date","application_date","inventors"
    ,"assignees","abstract","claims","description","drawings","similar_documents"
    ,"related_documents"],
  "size": 5
}
```

Рис. 4. – Запрос к системе ElasticSearch

```
"took": 97,
"timed_out": false,
"_shards": {
  "total": 1,
  "successful": 1,
  "skipped": 0,
  "failed": 0
},
"hits": {
  "total": {
    "value": 3,
    "relation": "eq"
  },
  "max_score": 2.421914,
```

Рис. 5. – Результат поиска в системе ElasticSearch

Выводы: Elasticsearch позволяет составлять поисковые запросы более одного слова как к одной части документа, так и к нескольким, что, в свою очередь обеспечивает очень удобный и точный настраиваемый поиск по базе данных.

Загрузим и запустим систему ***ClickHouse*** в Windows с помощью программы Docker. Проиндексируем патенты, чтобы мы могли добавить их в базу данных. Для поиска патентов в базе данных, используем следующий запрос: «медицинская техника»

В итоге на тестовой выборке было найдено 2 патента (Рис. 6).

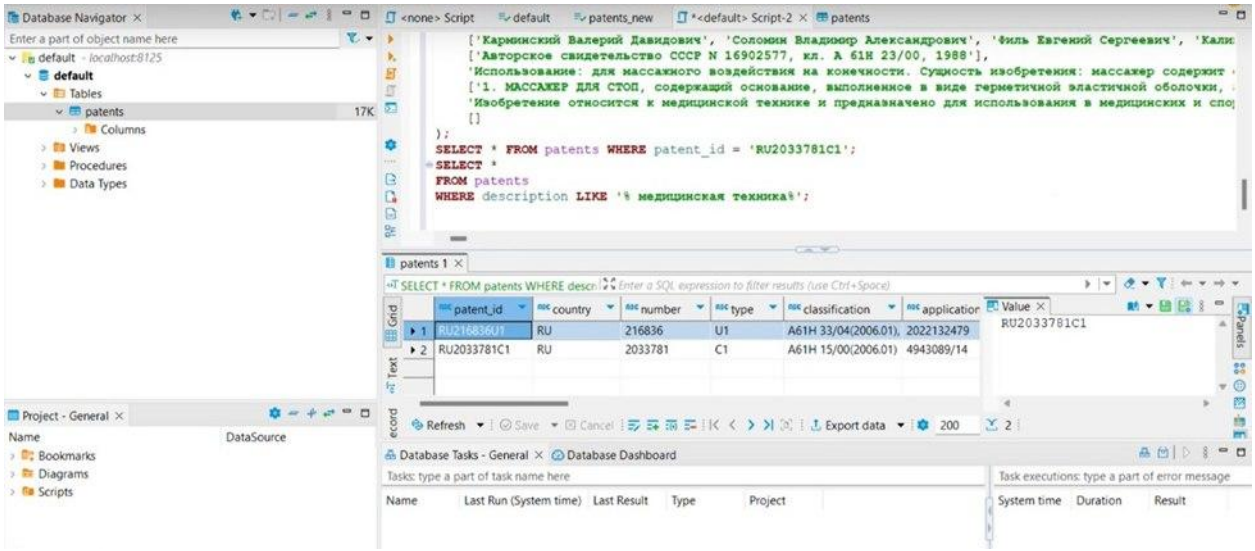

Рис. 6. – Запрос к системе ClickHouse

Результаты

Solr

Плюсы системы:

– Поддержка различных форматов данных (XML, JSON, CSV) и возможность настройки индексирования и поиска.

– Легко интегрируется с другими системами через RESTful API.

– Возможность фильтрации результатов по различным критериям.

– Может обрабатывать большие объемы данных с помощью распределенного поиска и репликации данных.

Минусы системы:

– Для достижения оптимальной производительности и функциональности требуется тщательная настройка.

– Производительность может быть ниже по сравнению с ElasticSearch при больших объемах данных и высоких нагрузках.

ElasticSearch

Плюсы системы:

– Elasticsearch позволяет составлять текстовые запросы (более одного слова) как к одной части документа, так и к нескольким, что, в свою очередь обеспечивает очень удобный и точный настраиваемый поиск по базе данных.

– Удобный интерфейс Kibana позволяет быстрее вводить запросы, тем самым делая работу в Elasticsearch эффективнее.

Минусы системы:

– Слова в разных падежах на русском языке не выводятся в запросах без предварительной установки hunspell и настройки индекса.

– При поиске нескольких слов в тексте нельзя проконтролировать расстояние между этими словами.

ClickHouse

Плюсы системы:

– Эффективное хранение и обработка данных для аналитических задач.

– Возможность выполнения сложных агрегаций и аналитики в реальном времени.

Минусы системы:

– Система не оптимизирована для полнотекстового поиска, в отличие от Solr и ElasticSearch.

– Требуется глубокое знание системы для оптимальной настройки и управления.

– Ограниченные возможности для настройки поиска и индексирования по сравнению с Solr и ElasticSearch.

Заключение

Проведен анализ системы полнотекстового поиска Solr. Рассмотрены архитектура Solr, основные функциональные возможности и особенности использования этой системы для поиска схожих текстовых документов. Были выявлены сильные стороны Solr, такие, как гибкость настройки схемы и возможность масштабирования через распределенный поиск.

Изучена и проанализирована ElasticSearch — ещё одна мощная система полнотекстового поиска. Рассмотрены её архитектура, основные функции и возможности, включая поддержку сложных запросов и высокую производительность. Проведенный анализ показал, что ElasticSearch предоставляет широкие возможности для эффективного поиска и анализа текстовых данных, особенно в условиях больших объемов данных.

Проанализирована система ClickHouse, которая, хотя и не является традиционной системой полнотекстового поиска, предлагает уникальные возможности для анализа больших объемов данных. Рассмотрены её архитектурные особенности, основные функции и методы обработки запросов. Анализ показал, что ClickHouse может быть эффективным инструментом для специфических задач, связанных с поиском и аналитикой текстовых данных.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 24-21-20140, rscf.ru/project/24-21-20140/, и Администрации Волгоградской области.

Литература (References)

1. Bobunov A., Korobkin D., Fomenkov S. Development of the Concept and Architecture of an Automated System for Updating Physical Knowledge for Information Support of Search Design. In 2023 International Russian Smart

Industry Conference, Sochi, Russian Federation. 2023. pp. 281-288. DOI: 10.1109/SmartIndustryCon57312.2023.10110764.

2. Korobkin D., Fomenkov S., Kravets A., Kolesnikov S. Prior art candidate search on base of statistical and semantic patent analysis. In Multi Conference on Computer Science and Information Systems; IADIS (International Association for Development of the Information Society), Lisbon, Portugal. 2017. pp. 231-238.

3. Vohra D. Apache Solr. In: Practical Hadoop Ecosystem. Apress, Berkeley, CA. 2016. DOI: 10.1007/978-1-4842-2199-0_10.

4. Ayyagiri A., Goel O., Jain S. Innovative Approaches to Full-Text Search with Solr and Lucene. Innovative Research Thoughts. 2024. 10 (3). pp. 144-159. DOI: 10.36676/irt.v10.i3.1473.

5. Hunt P., Konar M., Junqueira F., Reed B. ZooKeeper: Wait-free Coordination for Internet-scale Systems. USENIX annual technical conference. 2010. vol. 8. pp. 1-14.

6. Chen D., Chen Y., Brownlow B., Kanjamala P., Arredondo C., Radspinner B., Raveling M. Real-Time or Near Real-Time Persisting Daily Healthcare Data into HDFS and ElasticSearch Index inside a Big Data Platform. IEEE Transactions on Industrial Informatics. 2016. pp. 595-606. DOI: 10.1109/TII.2016.2645606.

7. Vohra D. Using Elasticsearch. In: Pro Couchbase Development. Apress, Berkeley, CA. 2015. DOI: 10.1007/978-1-4842-1434-3_7.

8. Prutzkow A. An Approach to Identify Books for the Initial Learning of the Elastic Stack and its Results. International Journal of Open Information Technologies. 2023. vol. 11. no. 11. pp. 53-57.

9. Bajer M. Building an IoT Data Hub with Elasticsearch, Logstash and Kibana. In 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), Prague, Czech Republic. 2017. pp. 63-682017. DOI: 10.1109/FiCloudW.2017.101.

10. Imasheva B., Nakispekov A., Sidelkovskaya A., Sidelkovskiy A. The Practice of Moving to Big Data on the Case of the NoSQL Database, Clickhouse. In: WCGO 2019. Advances in Intelligent Systems and Computing. 2020. vol. 991. DOI: 10.1007/978-3-030-21803-4_82.

11. Boyarsky J., Selikoff S. Welcome to Java. 2019. DOI: 10.1002/9781119584773.ch1.

12. Al-Hussaini L. Experience: Insights into the Benchmarking Data of Hunspell and Aspell Spell Checkers. Journal of Data and Information Quality. 2017. 8 (3-4). pp. 1-10. DOI: 1-10. 10.1145/3092700.

Дата поступления: 1.10.2024

Дата публикации: 17.11.2024