Оглавление

Введение

Изменения в различных сферах жизни общества происходят перманентно, в связи с этим меняется и законодательство, регулирующее эти сферы. Из-за большого числа нормативных актов и актов местного значения, сменяющих или дополняющих друг друга, возрастает потребность в автоматизации поиска, обработки и анализа законодательной базы. Ввиду того, что решение любой задачи на предприятии начинается с создания и анализа нормативной базы, существуют различные системы, обеспечивающие процесс поиска по нормативной базе на предприятии. Однако эти системы не способны удовлетворить запросы пользователей в полном объеме по причине разнообразия источников данных и их слабой формализации, процесс анализа нормативной базы осложняется необходимостью использования нескольких ресурсов.

В связи с этим возникает проблема организации поиска и анализа нормативной документации в рамках одного программного средства. Из данной проблемы следует тема работы “Разработка оболочки поисково-аналитической системы обеспечения правовой поддержки деятельности в области информационных технологий органов государственной и муниципальной власти на примере Министерства информационного развития и связи Пермского края”.

Цель исследования: разработка системы индексирования слабо формализованных данных для поисково-аналитической системы с тематическим поиском по всей базе нормативных документов, связанных с правовой деятельностью органов государственной и муниципальной власти.

Объект исследования: информация, в том числе слабо формализованная, ее хранение и обработка.

Предмет исследования: процессы индексирования и формирования результирующей выборки информации по запросу.

Гипотеза исследования. Результаты поиска будут релевантны запросу пользователя, содержать полный перечень документации по определенной

правовой теме, указанной в запросе, отображать связи между документами, если поиск организовать по алгоритму:

1) Анализ входных данных: определение источников (базы данных, почтовые серверы, сайты органов государственной и муниципальной власти, локальные директории на компьютерах сотрудников организации), выявление параметров документов (формат документов, их структура, определение внешних и внутренних ссылок);
2) Формирование индексов входных документов: автоматическое создание индексов средствами Solr, анализ полученных индексов и их обработка (переиндексирование данных с целью оптимизации поиска и хранения информации);
3) Организация взаимодействия клиентского приложения с поисковым сервером: обработка поисковых запросов, вывод результатов поиска.

На основании цели работы и гипотезы сформулированы следующие ***задачи***:

1. Проанализировать обеспечение правовой деятельности министерства информационного развития и связи Пермского края.
2. Изучить и проанализировать нормативно-правовую документацию, применяемую в осуществлении деятельности МИРСа.
3. Провести сравнительный анализ современных технологий, реализующих функцию полнотекстового поиска в массиве документов.
4. Создать полнотекстовые индексы для документов из нормативной базы (на примере МИРСа).
5. Разработать макет пользовательского интерфейса.
6. Выполнить тестирование разработанной системы индексации по тематическим запросам.

Теоретическая значимость: разработана теоретическая база исследования.

Практическая значимость: реализованы компоненты оболочки системы, объединение которых позволит обеспечить полнотекстовый поиск и анализ связей в нормативной базе предприятия.

Новизна:

— Разрабатываемая система реализует общий подход к индексации слабо формализованной информации, хранилище документов любой направленности и на основе сформированных индексов обеспечивает поисково-аналитическую обработку информации.

— Оболочка нацелена на тематический поиск, что позволит выдавать наиболее релевантные запросам пользователя результаты.

Апробация. Результаты работы представлены:

1. Статья:

Оценка информационной открытости Министерства информационного развития и связи Пермского края[1].

2. Организационная деятельность:

Координатор проекта “Оценка информационной открытости сайтов органов государственной и муниципальной власти”.

3. Участие в конференции:

Доклад «Оценка информационной открытости Министерства информационного развития и связи Пермского края» на совещании по обсуждению информационной открытости сайтов органов государственной власти Пермского края Регионального отделения Общероссийского общественного движения «Народный фронт «ЗА РОССИЮ» в Пермском крае.

Работа состоит из введения, трех глав, заключения, списка источников и четырех приложений.

Глава 1. Теоретические основы обработки

Законодательство представляет собой единую систему непротиворечивых взаимосвязанных нормативных документов. Каждый документ является лишь частью множества нормативных актов относящихся к определенной правовой проблеме. Для более глубокого анализа и принятия наиболее эффективного решения специалист должен изучить не только исходный документ, но и все документы, каким-либо образом связанные с ним.

В связи с этим возрастает потребность в создании информационно-поисковых систем, которые позволили бы организовать эффективную работу специалиста, анализирующего нормативную документацию в той или иной сфере. Такие системы должны обеспечить возможность полнотекстового поиска и навигации по связям между документами.

Информационно-поисковые системы (ИПС) предназначены для хранения и поиска информации, поиск в таких системах осуществляется на основе информационно-поискового языка (ИПЯ) и происходит по определенным правилам поиска. Данные в ИПС хранятся в специальной базе вместе с их описаниями (индексами), индексы позволяют системе быстро находить наиболее релевантные запросам пользователя данные.

1.1. Виды информационно-поисковых систем и их характеристики

В зависимости от типа, хранящегося в базе, объекта выделяют фактографические и документальные информационные системы (ИС).

Фактографические информационные системы (ФИС) обеспечивают хранение и поиск фактов. Фактографические системы оперируют сведениями, которые должны быть представлены в виде совокупности формализованной в записи данных. Такие системы используются для создания справочников в определенной предметной области, а также для

обработки хранящихся в системе данных. На рис. 1 представлена обобщенная схема фактографической информационной системы.

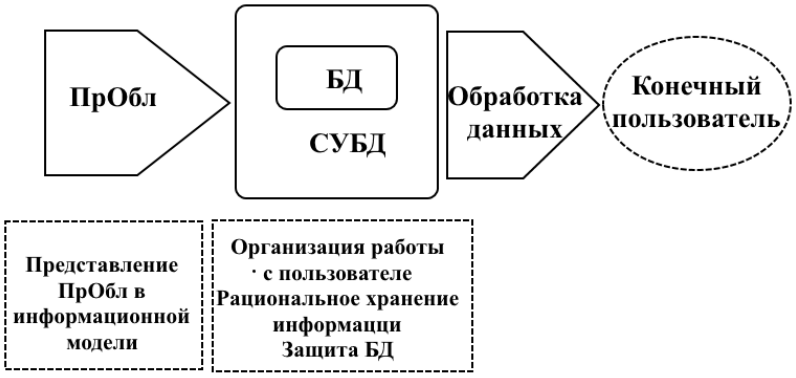

Рис. 1. Обобщенная схема ФИС

При создании информационной базы для фактографической системы входная информация должна пройти процесс структуризации. Объекты предметной области должны обладать свойствами, которые смогут их описать, значение одного и того же свойства может быть различным, но при этом оно должно выбираться из множества возможных значений (классификатора) или выражаться числом.

Документальные информационные системы (ДИС) предназначены для поиска по запросам пользователя в массиве документов и предоставления подмножества документов в качестве результата поиска. Обобщенная схема документальной ИС представлена на рисунке 2.

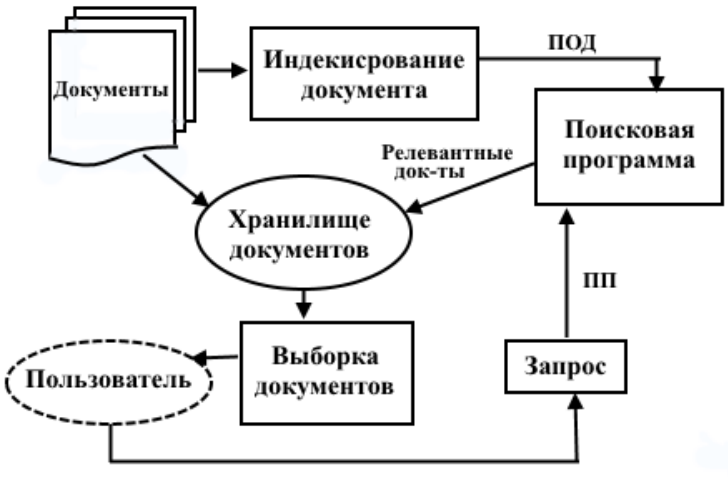

Рис. 2. Обобщенная схема ДИС

Для того чтобы автоматизировать процесс поиска по массиву документов, основное смысловое содержание поисковых запросов и документов формализуется в виде поисковых предписаний (ПП) и поисковых образов документов (ПОД) соответственно. Для записи ПП и ПОД применяются специальные информационно-поисковые языки. На основании набора правил об определении близости ПП и ПОД происходит отбор документов, которые отвечают потребностям пользователя.

1.1.1. Полнотекстовые ИПС

В полнотекстовых системах в процессе индексирования в индекс заносится информация обо всех словах в тексте документа. Такие системы строятся на основе ИПЯ дескрипторного типа. Структура полнотекстовых систем включает следующие элементы:

- хранилище документов;
- глобальный словарь системы;
- индекс документов инвертированного типа;
- интерфейс ввода документов в систему;
- механизм индексирования;
- механизм поиска документов;
- интерфейс запросов пользователя;
- механизм извлечения найденных документов.

В качестве дескрипторов ИПЯ системы выступают элементы глобального словаря. Документы, поступающие в систему, индексируются по глобальному словарю, при этом механизм индексирования полностью автоматизируется и заключается в создании двоичного вектора, указывающего наличие или отсутствие в данном документе слова с соответствующей позицией из глобального словаря. В результате индексирования поисковый образ каждого документа представляется набором словоформ из глобального словаря, которые присутствуют в тексте

документа, и поступает в виде соответствующего двоичного вектора для дополнения индекса системы. Индекс строится по инвертированной схеме и в двоичном виде отражает весь текст накопленных документов.

Пользователь формулирует свои информационные потребности по поиску документов путем создания поискового запроса, все запросы индексируются, как и документы в виде двоичного вектора поискового образа запроса, и передаются поисковой машине.

Поиск базируется на определенных алгоритмах и критериях сравнения поискового образа запроса с поисковыми образами документов. В результате определяются номера документов, которые соответствуют или близки поисковому запросу. Документы релевантные запросу пользователя формируются в выборку и передаются пользователю в качестве результата поиска.

1.1.1.1. Сравнение движков полнотекстового поиска

Поиск по тексту предполагает обработку больших объемов данных и сложных индексов, в связи с этим появляется множество отдельных инструментов, направленных на решение этой задачи.

В работу технологий полнотекстового поиска заложен следующий принцип: по текстовым данным формируется индекс, который обеспечивает быстрый поиск соответствий по ключевым словам.

Чаще всего поисковый сервис представлен двумя составляющими: индексатором и поисковиком. Индексатор, получив текст на вход, проводит его обработку – нормализацию слов, вырезание окончаний, отбрасывание незначительных слов и т.д. Обработка проводится в соответствии с указаниями схемы данных, которая содержит в себе информацию о том, каким образом индексатор должен преобразовать входной текст, и что должно обязательно остаться в индексе документа. Поисковик представляет собой интерфейс поиска по индексу, он принимает на вход запрос пользователя, обрабатывает его и ищет совпадения в индексе.

В настоящее время для реализации полнотекстового поиска существует несколько популярных технологий. В таблице 1 описаны их основные характеристики.

Таблица 1. Сравнительные характеристики поисковых движков

	MySQL	**Xapian**	**Sphinx**	**Solr**
Скорость индексации (МБ/с)	1,5	1,36	4,5	2,75
Скорость поиска средняя/ максимальная (мс)	175 / 3460	14 / 135	7 / 75	25 / 212
Размер индекса (% от размера данных)	150	200	30	20
Реализация	СУБД	Библиотека	Сервер	Сервер
Интерфейс	SQL	API	API, SQL	Web-сервис
Операторы поиска	Булевы, точная фраза. Префиксный поиск* (не рекомендуется использов	Булевы, префиксный поиск, точная фраза, слова вблизи, диапазоны, приблизительный поиск	Булевы, префиксный поиск, точная фраза, слова вблизи, диапазоны, порядок слов, зоны	Булевы, префиксный поиск + wildcard*, точная фраза, слова вблизи, диапазоны, приблизител

	ать по причине медленного отклика)			ьный поиск
Стеммеры*	-	15	15	31
Стоп-слова*, синонимы	-	+	+	+
Подсветка результатов	-	-	+	+
Дополнительные характеристики	Query Expansion (расширение запросов по словам из найденных документов)	Исправление опечаток, фасеты, “недовведенные” запросы	Синонимы со спецсимволами, зоны (абзацы/предложения/теги), слияние индексов	Гибкие настройки под требования пользователя, расширяемость

*Примечание к таблице 1:

Стеммер *(основа слова) - отбрасывает от слов окончания. Предназначен для поиска слов в различных падежах.*

Стоп-слова *— перечень часто употребляемых слов, индексация которых не имеет смысла, так как они не несут большого значения и встречаются почти везде.*

Префиксный поиск *— поиск слов, начинающихся на заданную последовательность символов. Wildcard-поиск — поиск слов, начинающихся на заданный префикс и оканчивающихся на заданный суффикс.*

Каждый из представленных в таблице движков имеет свои преимущества и недостатки. Наиболее подходящим для создания оболочки поисково-аналитической информационной системы является Solr, так как этот движок имеет множество полезных функций и способен к расширению, что позволит со временем полноценно использовать поисковые возможности в крупных организациях.

1.1.2. Гипертекстовые поисковые системы

Гипертекстовые системы позволяют искать информацию с учетом связей, имеющихся между документами, что делает процесс поиска эффективнее традиционных методов. Гипертекст организован таким образом, что текст представляет собой множество фрагментов с явно указанными ассоциативными связями между ними. Гипертекст можно представить как своеобразную базу данных, организованную в виде сети, узлы (фрагменты текста) которой соединяются пользователем.

Гипертекстовая информационно-поисковая система содержит в своей структуре следующие подсистемы:

- Подсистема отображения документов и ссылок. Отображение документа как в текстовом редакторе (документ сохраняет свою первоначальную структуру) и визуализация гиперссылок, содержащихся в этом документе.
- Подсистема навигации по гиперссылкам. Представляет собой интерфейс перехода по гиперссылкам. Система обеспечивает навигацию по связям документа путем скроллинга в случае, если ссылка указывает на фрагмент текста просматриваемого документа или открытием другого документа, если ссылка указывает на внешний источник. Чтобы обеспечить навигацию по гиперссылкам, для каждой связи хранится адрес расположения соответствующего фрагмента или файла.

- Подсистема формирования связей. Формирование связей может осуществляться двумя подходами – ручным или автоматизированным. Основным преимуществом ручного подхода является установление связей на основе многоаспектного анализа содержания документа, такой анализ не может быть проведен посредством автоматизированного алгоритма. Однако такой подход значительно уступает по производительности, а также требует определенной квалификации пользователя и глубокого знания предметной области. Автоматизированный подход применяется только в закрытых системах, и основывается на принципах поиска релевантных по смыслу документов.

- Хранилище документов. Существует несколько формальных моделей гипертекстовых структур, среди них теория паттернов У. Гренадера, тензорная модель А.В. Нестерова и логико-смысловое моделирование М.М. Субботина. В практическом применении наиболее часто используется логико-смысловое моделирование, которое позволяет связывать документы на основе семантической близости.

1.2. Индексация документов

Процесс индексирования базируется на совокупности правил и инструкций, включающих правила применения информационно-поискового языка. Под системой индексирования понимается комплекс методов и средств для перевода текстов с естественного языка на информационно-поисковый язык в соответствии с правилами применения ИПЯ и заданным набором словарей лексических единиц.

Классификация систем индексирования:

1. По степени автоматизации процесса индексирования:
    - системы ручного индексирования;

- системы автоматизированного индексирования;
- системы автоматического индексирования.

2. По степени контролируемости:
    - без словаря;
    - с жестким словарем;
    - со свободным словарем.
3. По характеру алгоритма по отбору слов из текста:
    - с последовательным просмотром текста (выбираются все полнозначные слова);
    - с эвристическими процедурами выбора слов (интуитивный отбор или по заданной процедуре);
    - со статистическими процедурами выбора слов (выбираются информативные слова в соответствии с распределением частот их употребления в тексте).
4. По характеру лексикографического контроля (устранение синонимии, полисемии, омонимии на основе нормативных словарей лексических единиц, а также приведение всех слов к нормальному виду с помощью морфологических нормативных словарей):
    - с полным контролем;
    - с промежуточным контролем;
    - без контроля.
5. По характеру морфологического анализа слов системы с использованием:
    - морфологических словарей;
    - основных лексических словарей;
    - морфологического анализа с усечение слов.

При индексации какого-либо текста с использованием систем с ручным индексированием индексатор выделяет слова и словосочетания, которые отражают содержание документа. Индексатор может

использовать слова, отсутствующие в тексте, но важные для выражения его смысла. Слова для формирования поискового образа документа индексатор может извлекать из индексируемого текста, словарей, энциклопедий и любых других доступных источников.

При использовании систем полусвободного индексирования процесс индексирования совпадает с вышеописанным процессом, но сформированный список слов сравнивается со словарем и несовпадающие слова исключаются из поискового образа документа.

В системах с жестким индексированием слова выбираются только из индексируемого текста, а в ПОД включаются только те слова, которые указаны в словаре. Перед формированием словаря для таких систем термины должны пройти процесс морфологической нормализации на основе лексических словарей.

В системах с автоматическим индексированием осуществляется последовательный автоматический поиск каждого ключевого слова в тексте документа. Таким образом строится индекс системы, реализующий поисковое пространство документов. Выделяют два типа образования индекса – прямой и инвертированный (рис. 3).

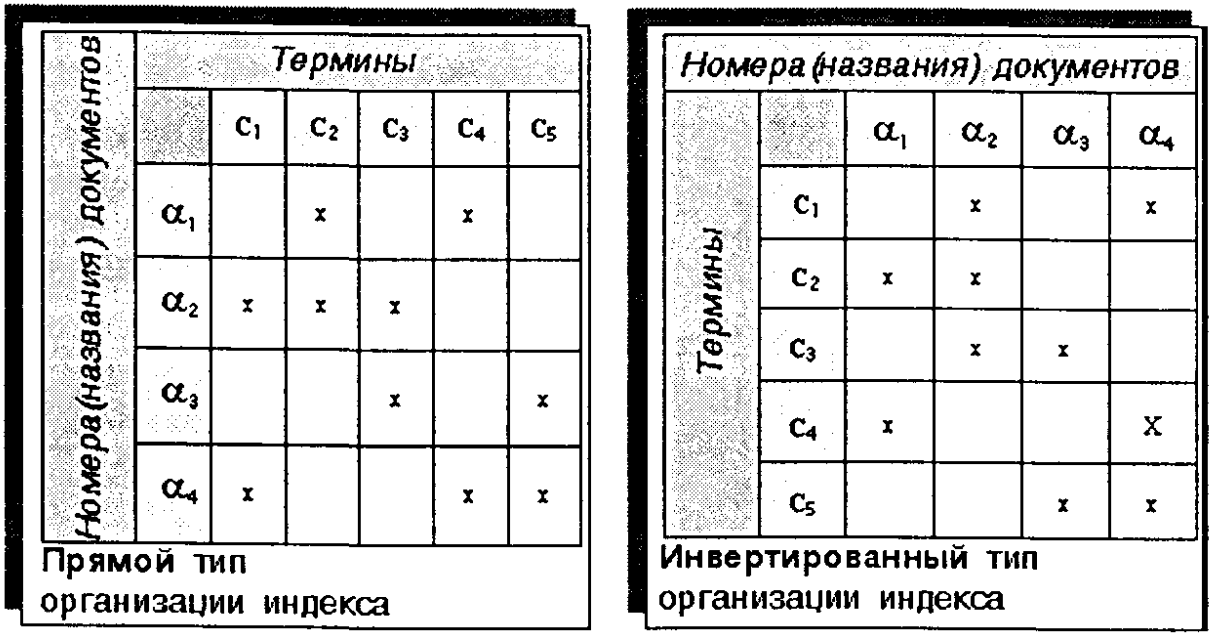

Рис. 3. Типы организации индекса[2]

Прямой тип индекса строится по схеме «Документ-термины», в этом случае поисковое пространство представлено в виде матрицы, где строки представляют поисковые образы документа. Инвертированный тип индекса строится по обратной схеме - «Термин-документы». Поисковое пространство представлено аналогичной матрицей, но в транспонированной форме, в этом случае поисковыми образами документов являются столбцы матрицы.

1.3. Частные задачи автоматизации отладки и тестирования поиска и определение места проектируемой подсистемы в ИС МИРСа

Министерство информационного развития и связи Пермского края обеспечивает выработку и реализацию региональной политики, и нормативно-правовое регулирование в сфере внедрения новейших информационных технологий в процесс предоставления государственных услуг, создания удобных интерактивных систем информационного обслуживания граждан, оказания услуг в области связи[3].

Организационная структура Министерства информационного развития и связи Пермского края представлена на рисунке №.

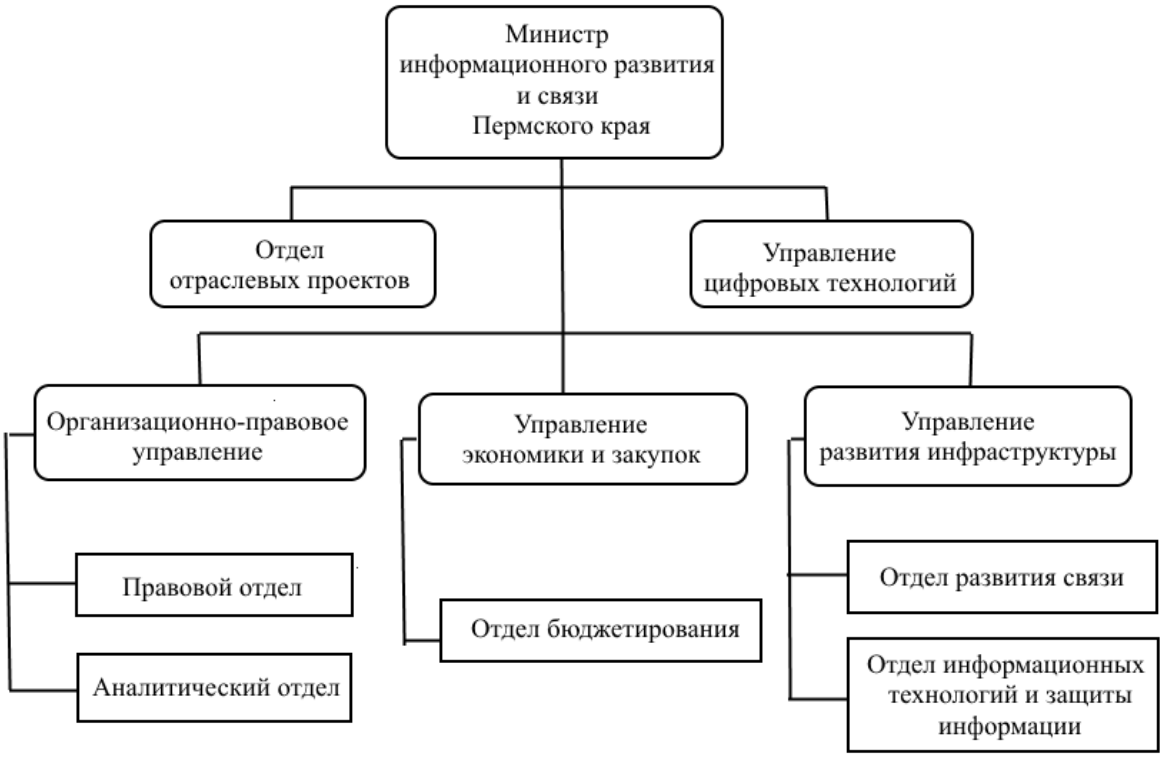

Рис. 4. Организационная структура МИРСа

Министерства информационного развития и связи Пермского края осуществляет следующие функции:

1. координирует взаимодействие органов государственной власти Пермского края, органов местного самоуправления муниципальных образований Пермского края, организаций по вопросам информатизации, связи и технической защиты информации, не содержащей сведения, составляющие государственную тайну;

2. осуществляет обеспечение развития интегрированных информационных систем и технологий, электронной связи, используемых в исполнительных органах государственной власти Пермского края, администрации губернатора Пермского края, аппарате Правительства Пермского края;

3. с участием заинтересованных органов государственной власти Пермского края, органов местного самоуправления муниципальных образований Пермского края и организаций координирует и организует работу по внедрению новейших технологий, созданию интерактивных

систем информационного обслуживания, оказанию новых услуг в области информационных технологий и связи;

4. обеспечивает надежность, устойчивость и безопасность функционирования информационно-коммуникационной инфраструктуры, используемой исполнительными органами государственной власти Пермского края, администрацией губернатора Пермского края, аппаратом Правительства Пермского края;

5. координирует процессы создания и развития информационной и телекоммуникационной инфраструктуры и связи Пермского края;

6. осуществляет совершенствование системы технической защиты информации, не содержащей сведения, составляющие государственную тайну, в исполнительных органах государственной власти Пермского края, администрации губернатора Пермского края, аппарате Правительства Пермского края;

7. осуществляет функции головного подразделения по технической защите информации, не содержащей сведения, составляющие государственную тайну, в Пермском крае;

8. формирует единую техническую политику, осуществляет организацию и координацию работ по технической защите информации, не содержащей сведения, составляющие государственную тайну, обрабатываемой в информационных системах исполнительных органов государственной власти Пермского края, администрации губернатора Пермского края, аппарата Правительства Пермского края;

9. осуществляет координацию деятельности подразделений (специалистов) по технической защите информации, не содержащей сведения, составляющие государственную тайну, в исполнительных органах государственной власти Пермского края, администрации губернатора Пермского края, аппарате Правительства Пермского края;

10. в пределах своей компетенции осуществляет функции оператора государственных информационных систем Пермского края в соответствии с требованиями законодательства Российской Федерации, Пермского края;

11. координирует предоставление исполнительными органами государственной власти Пермского края государственных услуг, оказываемых с использованием информационно-телекоммуникационных технологий;

12. координирует деятельность исполнительных органов государственной власти Пермского края по обеспечению информационной совместимости государственных, муниципальных и иных информационных систем в единой информационной системе межведомственного электронного взаимодействия в Пермском крае;

13. проводит экспертизу проектов административных регламентов предоставления государственных услуг в соответствии с Правиламипроведения экспертизы проектов административных регламентов предоставления государственных услуг, утвержденными Постановлением Правительства Пермского края от 8 мая 2013 г. № 417-п "О разработке административных регламентов предоставления государственных услуг и административных регламентов исполнения государственных функций, а также об экспертизе проектов административных регламентов предоставления государственных услуг";

14. организует работы по внедрению электронных подписей в деятельность исполнительных органов государственной власти Пермского края, администрации губернатора Пермского края, аппарата Правительства Пермского края, а также осуществляет взаимодействие с органами местного самоуправления муниципальных образований Пермского края по вопросам внедрения в их деятельность электронных подписей в соответствии с законодательством Российской Федерации;

15. обеспечивает внедрение и функционирование инструментов электронного Правительства Пермского края, Открытого Правительства Пермского края;

16. обеспечивает доступ к информации о деятельности Правительства Пермского края на официальном сайте Правительства Пермского края в информационно-телекоммуникационной сети "Интернет" в случаях и в порядке, предусмотренных законодательством, а также защиту указанной информации от неправомерных доступа, уничтожения, модифицирования, блокирования, копирования, предоставления, распространения и иных неправомерных действий;

17. организует информационное наполнение и техническое сопровождение официального сайта Правительства Пермского края в информационно-телекоммуникационной сети "Интернет" в государственной информационной системе "Портал Правительства Пермского края";

18. проводит анализ и мониторинг статистических данных о состоянии реализации региональной политики в сфере информатизации, тенденциях развития и использования информационных технологий, в том числе организует работы по проведению социологических исследований по оценке удовлетворенности граждан качеством предоставления государственных и муниципальных услуг, предоставляемых в соответствии с требованиями Федерального закона от 27 июля 2010 г. № 210-ФЗ "Об организации предоставления государственных и муниципальных услуг";

19. участвует в реализации федеральных, краевых целевых программ в пределах своей компетенции;

20. разрабатывает прогнозные показатели развития информатизации и связи в Пермском крае;

21. разрабатывает в установленном порядке программы развития информационных технологий в Пермском крае;

22. содействует организациям связи, оказывающим универсальные услуги связи, в получении и (или) строительстве сооружений связи и помещений, предназначенных для оказания универсальных услуг связи;

23. участвует в развитии и расширении сети почтовой связи, а также согласовании режима работы объектов почтовой связи организаций федеральной почтовой связи на территории Пермского края;

24. содействует операторам почтовой связи в расширении сферы услуг, предоставляемых гражданам и юридическим лицам;

25. вносит в федеральный орган исполнительной власти, осуществляющий управление деятельностью в области почтовой связи, предложения о совершенствовании и развитии сети почтовой связи на территории Пермского края;

26. проводит экспертизу проектов документов в сфере регулирования Министерства, готовит заключения, в том числе экспертные, методические материалы, аналитическую, справочную и иную информацию;

27. участвует в разработке технических средств и материалов по информатизации деятельности исполнительных органов государственной власти Пермского края, технической защите информации, не содержащей сведения, составляющие государственную тайну, проводит экспертизу представленных материалов;

28. осуществляет взаимодействие с Министерством экономического развития Российской Федерации, Министерством связи и массовых коммуникаций Российской Федерации по вопросам реализации Федерального закона от 27 июля 2010 г. № 210-ФЗ "Об организации предоставления государственных и муниципальных услуг", в том числе готовит в пределах своей компетенции отчеты и информацию по запросам указанных органов;

29. осуществляет функции по взаимодействию с краевыми организациями отраслевых профсоюзов;

30. организует в пределах своей компетенции прием граждан, юридических лиц, обеспечивает своевременное и полное рассмотрение устных и письменных обращений граждан, юридических лиц, принятие по ним решений и направление заявителям ответов в установленный законодательством срок;

31. осуществляет функции главного распорядителя средств бюджета Пермского края, функции администратора доходов бюджета Пермского края по вопросам ведения Министерства;

32. осуществляет функции по организации деятельности многофункциональных центров предоставления государственных и муниципальных услуг в соответствии с Федеральным законом от 27 июля 2010 г. № 210-ФЗ "Об организации предоставления государственных и муниципальных услуг";

33. в пределах установленной компетенции осуществляет функции учредителя (отраслевого органа) подведомственного краевого государственного автономного учреждения "Пермский краевой многофункциональный центр предоставления государственных и муниципальных услуг" (далее - подведомственное учреждение);

34. принимает решения о реорганизации и ликвидации подведомственного государственного учреждения Пермского края в порядке, установленном законодательством Российской Федерации и Пермского края;

35. осуществляет в пределах установленной компетенции функции отраслевого органа по управлению и распоряжению краевым имуществом подведомственного учреждения;

36. осуществляет финансовый контроль за подведомственным получателем бюджетных средств в части обеспечения правомерного, целевого и эффективного использования бюджетных средств;

37. в части осуществления прав акционера в отношении акционерных (хозяйственных) обществ, находящихся в управлении Министерства, Министерство;

38. организует и обеспечивает мобилизационную подготовку и мобилизацию Министерства, а также координирует и контролирует проведение организациями, которые находятся в сфере ведения Министерства, мероприятий по мобилизационной подготовке;

39. в пределах своей компетенции осуществляет планирование и организует проведение мероприятий по гражданской обороне в установленной сфере деятельности, в том числе участвует в разработке плана гражданской обороны и защиты населения Пермского края;

40. при введении гражданской обороны организует проведение мероприятий по координации деятельности организаций, предоставляющих услуги связи;

41. в пределах своей компетенции обеспечивает своевременное оповещение населения, в том числе экстренное оповещение населения, об опасностях, возникающих при ведении военных действий или вследствие этих действий, а также об угрозе возникновения или о возникновении чрезвычайных ситуаций природного и техногенного характера

42. обеспечивает в пределах своей компетенции защиту сведений, составляющих государственную тайну;

43. осуществляет иные функции в установленной сфере деятельности, предусмотренные действующим законодательством.

Полномочия Министерства информационного развития и связи Пермского края определяются следующим перечнем нормативных правовых актов:

- Федеральный закон от 24 июля 1998 г. № 124-ФЗ "Об основных гарантиях прав ребенка в Российской Федерации";
- Федеральный закон от 6 октября 1999 г. № 184-ФЗ "Об общих принципах организации законодательных (представительных) и исполнительных органов государственной власти субъектов Российской Федерации";
- Федеральный закон от 7 июля 2003 г. № 126-ФЗ "О связи";
- Федеральный закон от 27 июля 2006 г. № 149-ФЗ "Об информации, информационных технологиях и защите информации";
- Федеральный закон от 27 июля 2006 г. № 152-ФЗ "О персональных данных";
- Федеральный закон от 9 февраля 2009 г. № 8-ФЗ "Об обеспечении доступа к информации о деятельности государственных органов и органов местного самоуправления";
- Федеральный закон от 27 июля 2010 г. № 210-ФЗ "Об организации предоставления государственных и муниципальных услуг";
- Федеральный закон от 29 декабря 2010 г. № 436-ФЗ "О защите детей от информации, причиняющей вред их здоровью и развитию";
- Федеральный закон от 6 апреля 2011 г. № 63-ФЗ "Об электронной подписи";
- Федеральный закон от 28 июля 2012 г. № 133-ФЗ "О внесении изменений в отдельные законодательные акты Российской Федерации в целях устранения ограничений для предоставления государственных и муниципальных услуг по принципу "одного окна";

- Федеральный закон от 29 декабря 2012 г. № 273-ФЗ "Об образовании в Российской Федерации";
- Федеральный закон от 7 июня 2013 г. № 112-ФЗ "О внесении изменений в Федеральный закон "Об информации, информационных технологиях и о защите информации" и Федеральный закон "Об обеспечении доступа к информации о деятельности государственных органов и органов местного самоуправления";
- Стратегия развития информационного общества в Российской Федерации, утвержденная Президентом Российской Федерации 7 февраля 2008 г. № Пр-212;
- Указ Президента Российской Федерации от 7 мая 2012 г. № 601 "Об основных направлениях совершенствования системы государственного управления";
- Указ Президента Российской Федерации от 7 мая 2012 г. № 599 "О мерах по реализации государственной политики в области образования и науки";
- Указ Президента Российской Федерации от 1 июня 2012 г. № 761 "О Национальной стратегии действий в интересах детей на 2012-2017 годы";
- Указ Президента Российской Федерации от 10 сентября 2012 г. № 1274 "О Координационном совете при Президенте Российской Федерации по реализации Национальной стратегии действий в интересах детей на 2012-2017 годы";
- Указ Президента Российской Федерации от 4 марта 2013 г. № 183 "О рассмотрении общественных инициатив, направленных гражданами Российской Федерации с использованием интернет-ресурса "Российская общественная инициатива";

- Постановление Правительства Российской Федерации от 21 апреля 2005 г. № 241 "О мерах по организации оказания универсальных услуг связи";
- Постановление Правительства Российской Федерации от 8 сентября 2010 г. № 697 "О единой системе межведомственного электронного взаимодействия";
- Постановление Правительства Российской Федерации от 24 октября 2011 г. № 861 "О федеральных государственных информационных системах, обеспечивающих предоставление в электронной форме государственных и муниципальных услуг (осуществление функций)";
- Постановление Правительства Российской Федерации от 22 декабря 2012 г. № 1376 "Об утверждении Правил организации деятельности многофункциональных центров предоставления государственных и муниципальных услуг";
- Постановление Правительства Российской Федерации от 10 июля 2013 г. № 583 "Об обеспечении доступа к общедоступной информации о деятельности государственных органов и органов местного самоуправления в информационно-телекоммуникационной сети "Интернет" в форме открытых данных";
- Распоряжение Правительства Российской Федерации от 6 мая 2008 г. № 632-р "О Концепции формирования в Российской Федерации электронного правительства до 2010 года";
- Распоряжение Правительства Российской Федерации от 17 ноября 2008 г. № 1662-р "О Концепции долгосрочного социально-экономического развития Российской Федерации на период до 2020 года";
- Распоряжение Правительства Российской Федерации от 17 октября 2009 г. № 1555-р "О плане перехода на предоставление

государственных услуг и исполнение государственных функций в электронном виде федеральными органами исполнительной власти";

- Распоряжение Правительства Российской Федерации от 30 декабря 2013 г. № 2602-р "Об утверждении плана мероприятий ("дорожной карты") "Развитие отрасли информационных технологий";
- Распоряжение Правительства Российской Федерации от 25 апреля 2011 г. № 729-р "Об утверждении перечня услуг, оказываемых государственными и муниципальными учреждениями и другими организациями, в которых размещается государственное задание (заказ) или муниципальное задание (заказ), подлежащих включению в реестры государственных или муниципальных услуг и предоставляемых в электронной форме";
- Распоряжение Правительства Российской Федерации от 28 декабря 2011 г. № 2415-р "О государственных и муниципальных услугах, предоставляемых в электронном виде";
- Распоряжение Правительства Российской Федерации от 17 декабря 2012 г. № 1993-р "Об утверждении сводного перечня первоочередных государственных и муниципальных услуг, предоставляемых органами исполнительной власти субъектов Российской Федерации и органами местного самоуправления в электронном виде, а также услуг, предоставляемых в электронном виде учреждениями и организациями субъектов Российской Федерации и муниципальными учреждениями и организациями";
- Распоряжение Правительства Российской Федерации от 10 июля 2013 г. № 1187-р "О Перечнях информации о деятельности государственных органов, органов местного самоуправления, размещаемой в сети "Интернет" в форме открытых данных";
- Постановление Правительства Российской Федерации от 15 апреля 2014 г. № 313 "Об утверждении государственной программы

Российской Федерации "Информационное общество (2011-2020 годы)";

- Закон Пермского края от 2 апреля 2010 г. № 598-ПК "О стратегическом планировании социально-экономического развития Пермского края";
- Закон Пермского края от 20 декабря 2012 г. № 140-ПК "О Программе социально-экономического развития Пермского края на 2012-2016 годы" (далее - Программа социально-экономического развития Пермского края);
- Постановление Законодательного Собрания Пермского края от 1 декабря 2011 г. № 3046 "О Стратегии социально-экономического развития Пермского края до 2026 года";
- Указ губернатора Пермского края от 27 июля 2011 г. № 62 "Об утверждении Положения о планировании и реализации мероприятий, связанных с применением информационных технологий при создании, эксплуатации и модернизации автоматизированных информационных систем и их отдельных компонентов в исполнительных органах государственной власти Пермского края, администрации губернатора Пермского края и аппарате Правительства Пермского края";
- Указ губернатора Пермского края от 18 октября 2012 г. № 82 "Об утверждении Положения о координационном совете по реализации Национальной стратегии действий в интересах детей на 2012-2017 годы";
- Указ губернатора Пермского края от 24 мая 2013 г. № 60 "Об утверждении Региональной стратегии действий в интересах детей в Пермском крае на 2013-2017 годы";
- Постановление Правительства Пермского края от 10 января 2012 г. № 10-п "Об утверждении Перечня услуг, которые являются

необходимыми и обязательными для предоставления исполнительными органами государственной власти Пермского края государственных услуг, и Порядка определения размера платы за оказание услуг, которые являются необходимыми и обязательными для предоставления исполнительными органами государственной власти Пермского края государственных услуг";

- Постановление Правительства Пермского края от 10 января 2012 г. № 2-п "Об утверждении Перечня государственных услуг, предоставление которых организуется в краевом государственном автономном учреждении "Пермский краевой многофункциональный центр";
- Постановление Правительства Пермского края от 6 июля 2012 г. № 486-п "Об организации межведомственного информационного взаимодействия при предоставлении государственных услуг исполнительными органами государственной власти Пермского края и муниципальных услуг";
- Постановление Правительства Пермского края от 15 апреля 2013 г. № 255-п "Об утверждении Положения об особенностях подачи и рассмотрения жалоб на решения и действия (бездействие) исполнительных органов государственной власти Пермского края и их должностных лиц, государственных гражданских служащих Пермского края";
- Постановление Правительства Пермского края от 8 мая 2013 г. № 417-п "О разработке административных регламентов предоставления государственных услуг и административных регламентов исполнения государственных функций, а также об экспертизе проектов административных регламентов предоставления государственных услуг";

- Распоряжение Правительства Пермского края от 31 октября 2011 г. № 205-рп "О создании краевого государственного автономного учреждения "Пермский краевой многофункциональный центр".

На данный момент в Министерстве информационного развития и связи Пермского края для управления правовой информацией используется информационная система “Система электронного документооборота” (СЭД). Данная система позволяет создавать, хранить, просматривать нормативные акты, связанные с деятельностью министерства, распространять правовую информацию среди пользователей данной системы, осуществлять поиск необходимых документов по присвоенному системой номеру. Также система содержит все необходимые шаблоны для создания внутриведомственных нормативных актов, таких как приказы, распоряжения, письма и т.д. Однако СЭД имеет ряд недостатков. Разграничение прав доступа работает нестабильно, что может существенно замедлить работу с правовой информацией. Заявленный функционал позволяет искать документ по его названию, но данная функция выдает некорректные результаты, поэтому пользователи вынуждены тратить время на набор длинных номеров СЭД. Нормативные документы федерального уровня в СЭД отсутствуют, для получения информации по этим документам сотрудники обращаются к системе КонсультантПлюс или другим источникам.

1.3.1. Определение комплекса задач автоматизации

В рамках осуществления своей деятельности Министерство информационного развития и связи опирается на базу нормативно-правовых актов, регламентирующих полномочия министерства. Всю базу нормативной документации, направленной на реализацию деятельности МИРСа, можно разделить по уровню органа государственной власти, создавшего тот или иной нормативный документ. Можно выделить следующие документы, необходимые для осуществления деятельности министерства:

— Федеральные законодательные акты (Конституция, федеральные законы, постановления правительства РФ, указы Президента РФ и прочие);
— Региональные законодательные акты (Постановления правительства Пермского края, указы губернатора, положения и т.д.);
— Локальные документы можно разделить на:
  - Нормативные (указы, положения, приказы и прочие)
  - Ненормативные (регламенты, технические задания и т.д.).

Ввиду большого объема нормативной документации, используемой при осуществлении деятельности Министерства информационного развития и связи, возрастает актуальность оптимизации процесса поиска, который позволит в ответ на запрос пользователя системы выдавать необходимые документы и не отображать документы, которые не соответствуют запросу.

Также, чтобы оптимизировать процесс поиска, нужно сформировать единый распределенный источник данных, который объединит в себе документы с различных сайтов, базы данных, почтовые серверы и данные, хранящиеся на локальных компьютерах пользователей. Это позволит извлекать необходимые документы в режиме одного окна. Для наиболее эффективной навигации по нормативно-правовой базе необходимо реализовать вывод списка документов, которые ссылаются на просматриваемый документ, и списка документов, на которые ссылается текущий документ.

Можно выделить следующие задачи, направленные на оптимизацию поиска:

- Сформировать массив входных документов, используя в качестве источников локальные директории пользовательских ПК, документы, размещенные в базах данных, документы, опубликованные на официальных сайтах государственных органов;

- Организовать процесс индексации входных данных для реализации полнотекстового поиска. Для этого необходимо провести настройку соответствующего программного обеспечения и сформировать базу данных для хранения реквизитов документов;
- Спроектировать и разработать подсистему навигации по связям между документами;
- Разработать пользовательский интерфейс для обеспечения информационных потребностей пользователей.

1.3.2. Место проектируемой задачи в комплексе задач и ее описание

Для определения места проектируемой задачи в комплексе задач были разработаны IDEF-диаграммы, описывающие процесс поиска информации в организации после оптимизации. Деятельность МИРСа представлена контекстной диаграммой (рис. 5).

Входными данными являются источники документов такие как: коллекции документов, хранящиеся на локальных компьютерах пользователей системы; документы, размещенные в СЭД; документы, опубликованные на официальных сайтах государственных органов (сайты Президента РФ, Правительства РФ, Правительства Пермского края, Администрации губернатора Пермского края и т.д.).

В качестве выходных параметров выступают выборки документов, релевантных запросам пользователей, и хранилище реквизитов исходных документов, по которым осуществляется поиск.

Рис. 5. Контекстная диаграмма Министерства информационного развития и связи

Система функционирует на основании действующего законодательства, регламентирующего процессы использования информации, создаваемой органами государственной власти и устава организации.

Пользователями системы являются сотрудники Министерства информационного развития и связи Пермского края. Пользователей можно разделить на две категории:

- Администратор. Данная категория обеспечивает процессы формирования источников данных, подготовки данных для полнотекстового индексирования и настройки программного обеспечения.
- Конечный пользователь. Эта категория пользователей формирует поисковые запросы для определения своих информационных потребностей и является получателем данных, отобранных системой в соответствии с запросом пользователя.

Для автоматизации деятельности организации в данной информационной системе рассматриваются бизнес-процессы, которые

необходимо автоматизировать. Диаграмма декомпозиции контекстной диаграммы представлена на рисунке 6.

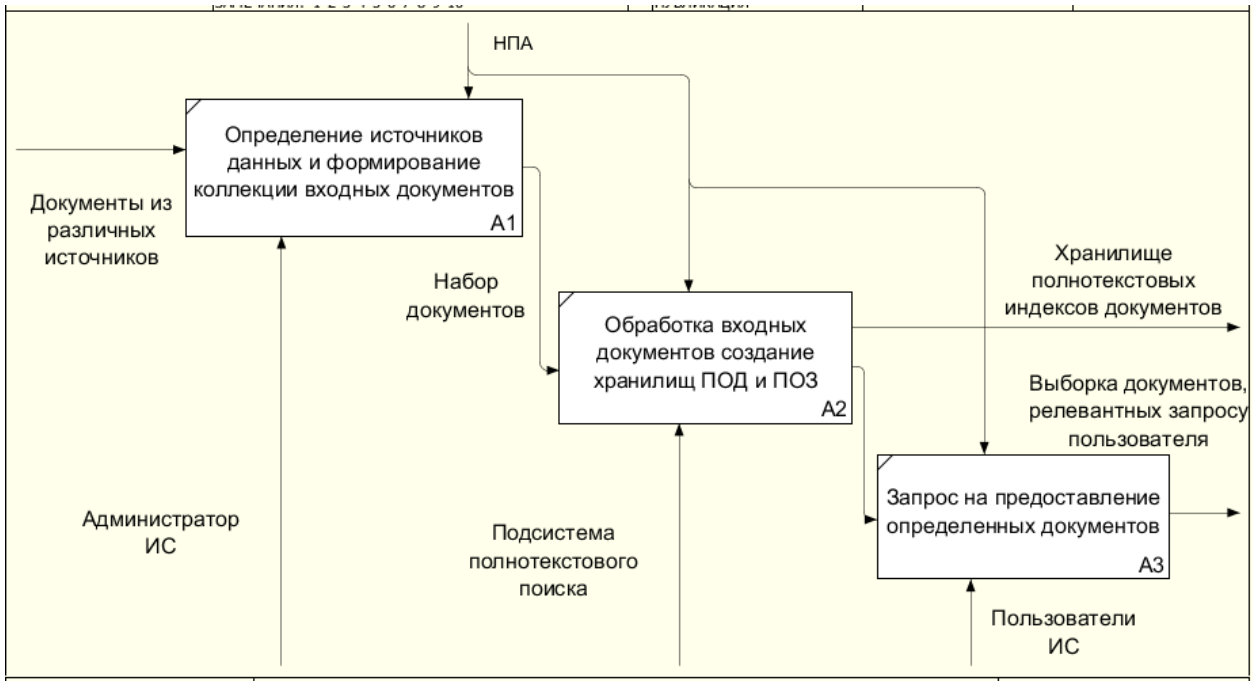

Рис. 6. Диаграмма декомпозиции контекстной диаграммы Министерства информационного развития и связи

В рамках декомпозиции были выделены следующие подпроцессы:

- Определение источников данных и формирование коллекции входных документов;
- Обработка входных документов, создание хранилищ для поисковых образов документов и запросов;
- Формирование запроса на предоставление определенных документов.

Входными данными для блока «Определение источников данных и формирование коллекции входных документов» являются документы из различных источников. На выходе формируется набор документов, по которому будет осуществляться поиск. Работу с системой на этом этапе проводит администратор системы.

На вход блока «Обработка входных документов, создание хранилищ для поисковых образов документов и запросов» подается набор документов,

сформированный на предыдущем этапе. В этом блоке входной набор данных проходит соответствующую обработку, необходимую для создания поисковых образов документов. Поисковые образы документов формируют хранилище, в котором в последующем будет проводиться процесс поиска по запросу пользователя. Работу этого блока осуществляет поисковая система при формировании специальных команд администратором системы.

Для блока «Формирование запроса на предоставление определенных документов» входными данными являются запросы пользователей на предоставление определенного набора документов. На выходе формируется выборка документов, соответствующих запросу пользователя. Работу с этим блоком осуществляют все сотрудники организации.

На всех этапах система функционирует в соответствии с действующим законодательством, регламентирующим процессы использования информации, создаваемой органами государственной власти, и уставом организации.

1.3.3. Спецификация требований к информационной системе

Оболочка поисково-аналитической системы направлена на поддержку правовой деятельности органов государственной и муниципальной власти путем оптимизации процесса поиска и анализа массивов нормативной документации, относящихся к той или иной сфере государственного управления.

Перед разработкой системы можно выделить следующие требования:

— Бизнес-требования:

- Выдача релевантных результатов в соответствии с информационным запросом пользователя;
- Формирование полнотекстовых индексов для нормативных документов;

- Оптимизация поиска необходимой документации.

— Пользовательские требования:

- Обработка информационных запросов;
- Отображение связей между документами;
- Формирование тематических подборок документов в соответствии с запросом.

— Функциональные и нефункциональные требования (Таблица 2).

Таблица 2. Основы программных требований.

Функциональные требования	Нефункциональные требования
• Система должна позволять производить полнотекстовый поиск по нормативной базе документов; • Система должна позволять просматривать подборки документов соответствующих запросу и анализировать связи между документами; • Система должна позволять эффективно взаимодействовать с массивом нормативной документации путем снижения временных затрат на поиск и анализ необходимой информации.	• Федеральный закон от 27 июля 2006 г. № 149-ФЗ "Об информации, информационных технологиях и о защите информации"; • Федеральный закон "Об обеспечении доступа к информации о деятельности государственных органов и органов местного самоуправления" от 09.02.2009 N 8-ФЗ; • ГОСТ 7.66-92 “Индексирование документов”.

1.3.4. Регламент работы оболочки в нотациях DFD

Для описания внешних по отношению к системе источников данных, потоков и хранилищ, к которым осуществляется доступ, используют диаграммы потоков. На рисунке 7 представлена диаграмма потоков для оболочки поисково-аналитической системы.

Перед началом работы системы формируется хранилище источников данных, в которых содержатся исходные документы. Данное хранилище представляет собой массив документов и списка их расположения.

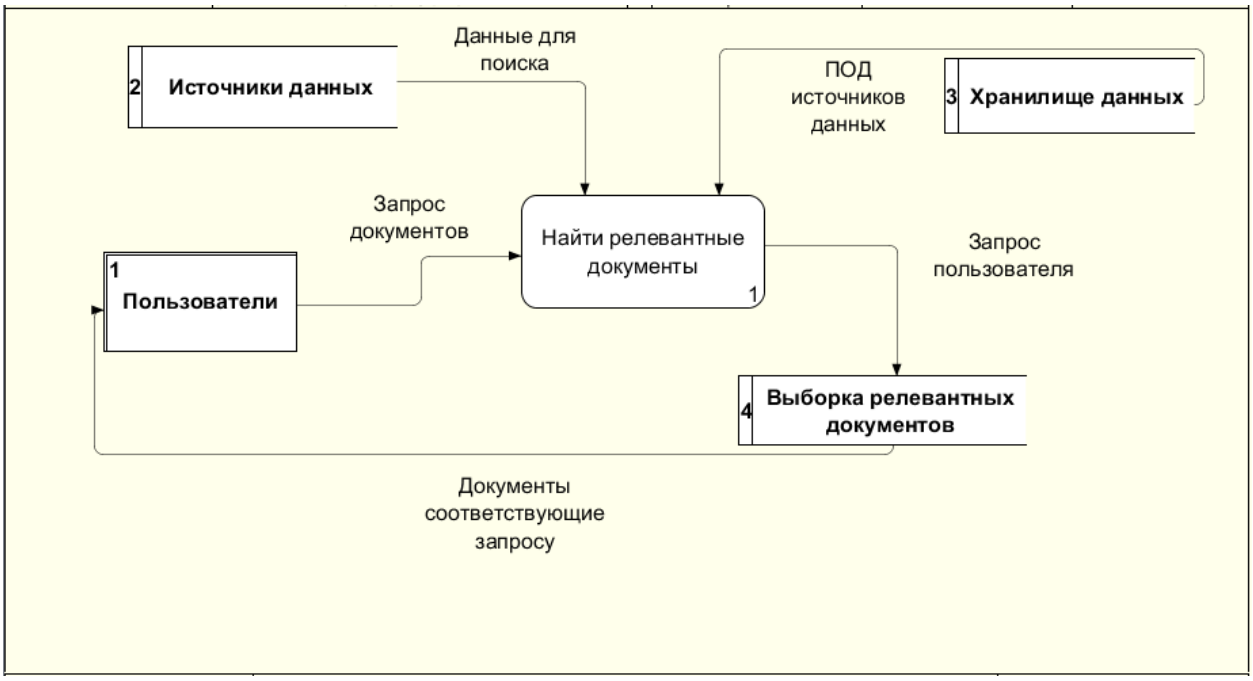

Рис. 7. Диаграмма потоков данных для оболочки поисково-аналитической системы

После проведения процесса индексации входных данных сформированные индексы документов образуют хранилище поисковых образов данных. При обработке запроса пользователя система производит поиск соответствий в хранилище данных и формирует выборку релевантной информации. Выборка релевантных документов содержит в себе результат обработки пользовательского запроса в виде указания на источники документов.

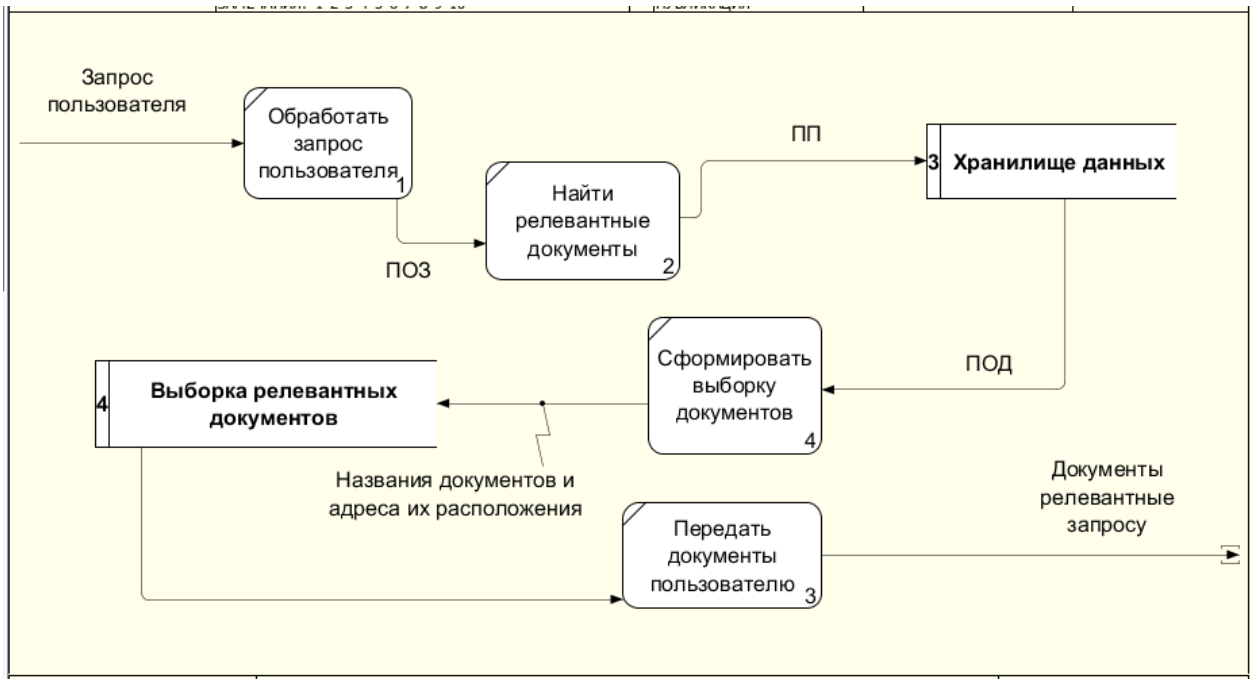

Рис. 8. Диаграмма регламента работы оболочки системы

Процесс взаимодействия пользователей с системой представлен на рисунке 8. Пользователь формулирует запрос и передает его системе, которая его обрабатывает. После обработки запроса система обращается к хранилищу данных, где производит поиск соответствий между поисковыми образами документов и поисковым образом запроса. Определив массив образов документов, которые соответствуют запросу, система формирует выборку релевантных документов. Выборка результатов поиска передается пользователю в виде списка документов и их расположения.

Выводы по первой главе

В связи с возрастающим объемом данных, преимущественно неформализованных, возникает необходимость в организации оптимального поиска и анализа накопленной информации. Наиболее подходящими средствами для решения этой задачи являются документальные информационно-поисковые системы с поддержкой функций полнотекстового поиска и анализа связей между данными. Однако на данный момент не

существует программных средств, комплексно реализующих обе эти функции. В данной главе были сформированы задачи по созданию программного продукта, отвечающего данному требованию:

— Определить массив источников данных;
— Настроить систему создания полнотекстовых индексов;
— Провести анализ входных документов с целью обнаружения связей между ними;
— Разработать пользовательский интерфейс для обработки информационных запросов.

В процессе сравнения существующих программных средств было выделено наиболее подходящее – Solr, как многофункциональный, масштабируемый и гибко настраиваемый поисковый движок.

В итоге, реализацию приложения в данной работе можно описать следующим алгоритмом:

1. Сформировать выборку входных документов.
2. Проанализировать документы с целью выявления связей между ними.
3. Реализовать установку и настройку поискового движка Solr для формирования полнотекстовых индексов документов.
4. Обработать автоматически созданные индексы для оптимизации хранения и поиска информации.
5. Спроектировать пользовательский интерфейс.

Глава 2. Система индексации и анализа слабоформализованных данных

Для реализации алгоритма по созданию программного продукта необходимо:

- Ознакомиться с принципами индексации документов;
- Изучить техническую документацию Solr;
- Настроить процесс автоматической индексации документов средствами Solr;
- Провести анализ сформированных индексов и переиндексировать данные с учетом добавления связей между документами;
- Спроектировать макет пользовательского интерфейса для взаимодействия с сервером Solr.

2.1. Индексирование документов средствами Solr

Solr поддерживает различные форматы входных данных, основными форматами, с которыми работает Apache Solr , являются XML, JSON и CSV. Также Solr поддерживает индексацию таких популярных форматов как pdf, rtf, форматы office (текстовые документы, таблицы, презентации и пр.). В зависимости от типов входных данных, нужно определить схему, которая указывает формат данных и методы их обработки.

Для обработки данных в процессе индексирования и поиска Solr разбивает текст на токены с помощью трех основных компонентов: анализаторов, токенизаторов и фильтров (Рис. 9).

Анализаторы - это основные компоненты, которые обрабатывают введенный текст при индексировании и поиске. Роль анализатора состоит в том, чтобы исследовать входной текст и сгенерировать токены.

Функция *токенизатора* состоит в том, чтобы разбить входной текст на токены, где каждый токен представляет собой поток символов в тексте.

Есть много классов, которые включены в релиз Solr. Ниже описаны некоторые из них:

- *Стандартный токенизатор* разбивает входной текст на токены, рассматривая пробелы и пунктуации как разделители. Это наиболее используемый токенизатор в конфигурации Solr.
- *Keyword tokenizer* рассматривает весь входной текст как один токен.
- *Lowercase tokenizer* токенизирует входной текст, преобразуя все буквы в строчные. В этом токенизаторе пробелы и символы отбрасываются.

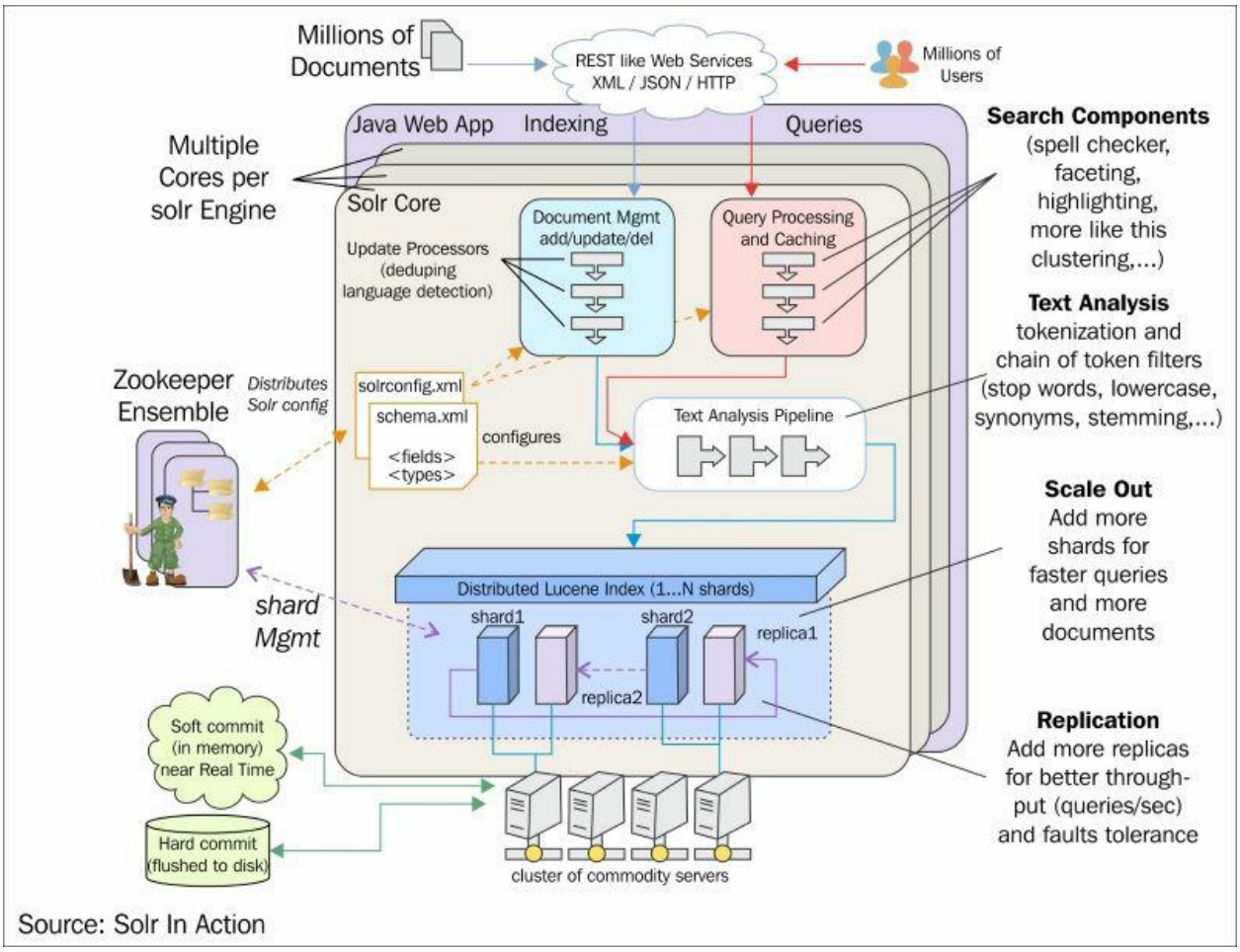

Рис. 9. Схема работы Solr[4]

Фильтры, как и токенизаторы, потребляют токены в качестве входных данных и снова создают поток токенов. Функция фильтра немного отличается от функции токенизатора. В отличие от токенизатора, фильтр принимает токены в качестве входа (передается токенизатором), а его

функция - смотреть на каждый токен и решать, хранить ли этот токен, изменять / заменять его или отбрасывать.

Наиболее распространенные фильтры:

- *Lowercase filter* преобразует все прописные буквы в строчные символы, а все остальные символы остаются неизменными.
- *Synonym filter* отвечает за сопоставление синонимов. Каждый токен сопоставляется со списком синонимов, присутствующих в файле синонимов, переданным в качестве аргумента, и если совпадение найдено, тогда синоним помещается вместо токена.

2.1.1. Техника и использование обработчиков индексов при индексировании данных

Существует много способов отправки данных в Solr с использованием API или путем вызова функции POST для использования инструмента post.jar или post.sh для отправки данных на сервер Solr для индексации.

Solr поставляется с большим количеством плагинов, которые можно использовать для импорта документов из большого количества источников. Документы можно индексировать с помощью Apache Tika – это позволит индексировать документы, таких форматов как MS Word, электронные таблицы Excel, документы PDF и многие другие форматы файлов.

Кроме того, Solr предоставляет возможность импорта данных из реляционных баз данных или структурированных типов данных с использованием обработчика импорта данных.

2.1.2. Индексирование данных с помощью Apache Tika и Apache Nutch

Apache Tika - это библиотека с открытым исходным кодом, которая используется для обнаружения типа документа и извлечения содержимого из

различных форматов файлов. Он использует различные существующие анализаторы документов и методы обнаружения типов документов для обнаружения и извлечения данных. Используя Tika, можно разработать детектор универсального типа и экстрактор содержимого для извлечения как структурированного текста, так и метаданных из разных типов документов, таких как электронные таблицы, текстовые документы, изображения, PDF-файлы и даже мультимедийные входные форматы.

Apache Nutch - это поисковый робот с открытым исходным кодом, который может использоваться для извлечения данных с веб-сайтов и получения данных из него. Это расширяемый и масштабируемый искатель, который дает нам свободу использовать его, как нам нравится, с помощью плагинов. Apache Nutch написан на Java, как и Apache Solr, и оба инструмента делают идеальную комбинацию для создания собственной поисковой системы, если они объединены.

2.2. Архитектура ПО для обеспечения информационного взаимодействия

Оболочка представляет собой совокупность модулей для взаимодействия с поисковым сервером Solr. Архитектура ПО для обеспечения информационного взаимодействия (рис. 10) содержит два модуля: модуль обработки запросов пользователей и модуль анализа индексов документа.

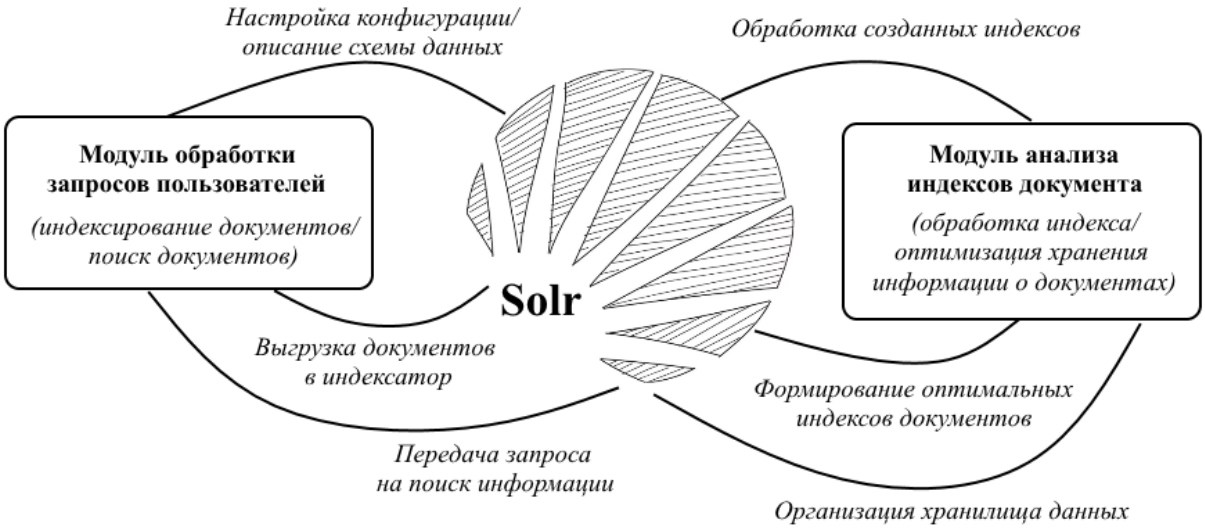

Рис. 10. Архитектура ПО для обеспечения информационного взаимодействия

Модуль обработки запросов пользователей выполняет следующие функции:

- Настройка конфигурации Apache Solr и описание схемы данных;
- Обработка запросов пользователей на индексирование входных данных;
- Выгрузка входной документации в индекс Solr;
- Обработка запросов пользователей на поиск информации в базе;
- Передача запроса на поиск информации поисковому серверу;
- Передача ответа на запрос от сервера клиенту.

Модуль анализа индексов документа выполняет следующие действия:

- Анализ автоматически созданных индексов;
- Определение оптимальной структуры индекса для оптимизации процесса поиска;
- Переиндексация данных с учетом оптимальной структуры индекса и необходимых реквизитов документа, отображаемых в индексе;
- Организация хранения обработанных индексов документов во встроенной базе Solr.

Организация архитектуры подобным образом позволяет оптимально взаимодействовать с пользователями системы и с входящими данными. Взаимосвязь модулей обработки информации обеспечивает возможность повысить релевантность поисковых результатов и в полном объеме удовлетворить информационные потребности пользователя.

2.3. Описание входных данных на примере нормативной базы МИРСа

Входные данные представлены массивом нормативно-правовой документации. Для каждого документа можно выделить реквизиты (таблица 3).

Таблица 3. Реквизиты входных документов

Поле	Тип данных	Описание
id	Строка	Идентификатор
author	Строка	Наименование гос органа, издавшего документ
date	Дата	Дата публикации
links	Массив строк	Идентификаторы документов, на которые ссылается документ
title	Строка	Название документа
lvl	Строка	Законодательный уровень документа (федеральный, региональный, локальный)
location	Строка	Расположение документа (путь к документу на компьютере/в базе/url ссылка)
status	Логическая переменная	Действующий/не действующий документ
content	Строка	Содержание документа
typedoc	Строка	Формат исходного документа

От первоначального индекса, созданного Solr (автоматически созданный индекс представлен в Приложении 2), в конечный индекс переносится поле “content”, остальные поля индекса определяются администратором. Solr по умолчанию в поле “id” устанавливается путь к файлу, после переиндексации в поле “id” вносится идентификатор файла, а путь в хранилище файлов указывается в поле “location”.

2.4. Описание процесса индексирования входных данных

Массив документов, хранящихся в директории локального компьютера, подается на вход обработчику документов. На основании конфигурации и схемы данных документы обрабатываются средствами Solr’s ExtractingRequestHandler. ExtractingRequestHandler использует Tika, чтобы позволить пользователям загружать двоичные файлы в Solr и извлекать из него текст, а затем его индексировать. Tika автоматически определяет тип входного документа (слово, pdf и т. д.) и соответствующим образом извлекает контент. Tika создает поток XHTML, который передается в SAX ContentHandler. Затем Solr реагирует на события SAX Tika и создает поля для индексации. Tika создает метаданные, такие как Title, Subject и Author, в соответствии со спецификациями, а извлеченный текст добавляется в поле “content”.

Для того, чтобы обеспечить индексирование файлов pdf, doc и др. форматов, необходимо настроить конфигурацию Solr следующим образом:

- Подключить библиотеку обработки двоичных файлов:

```
<lib dir="${solr.install.dir:../../../..}/contrib/extraction/lib" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-cell-\d.*\.jar" />
```

- Подключить плагин ExtractingRequestHandler:

```
<requestHandler name="/update/extract"
                startup="lazy"
```

```
            class="solr.extraction.ExtractingRequestHandler" >
    <lst name="defaults">
      <str name="xpath">/xhtml:html/xhtml:body/descendant:node()</str>
      <str name="capture">content</str>
      <str name="fmap.meta">attr_meta_</str>
      <str name="uprefix">attr_</str>
      <str name="lowernames">true</str>
    </lst>
  </requestHandler>
```

После обработки для каждого документа создается индекс. Для обеспечения наиболее релевантного поиска по сформированному индексу необходимо провести процесс адаптации индекса под параметры предметной области, т.е. обработать индекс таким образом, чтобы в нем отсутствовали лишние поля и присутствовали все необходимые для поиска данные.

С помощью эксперта для каждого документа выявляются нормативные связи между документами: внешние ссылки (документы, влияющие на текущий), внутренние ссылки (влияние текущего документа на другие нормативные акты). Например, документ “Постановление от 1 декабря 2011 г. N 3046 «О стратегии социально-экономического развития Пермского края до 2026 года»” имеет связи, представленные в таблице 4.

Таблица 4. Связи документа “Постановление от 1 декабря 2011 г. N 3046 «О стратегии социально-экономического развития Пермского края до 2026 года»”

Внешние ссылки	**Внутренние ссылки**
Постановление Законодательного Собрания Пермского края от 06.12.2012 N 569	Стратегия социально-экономического развития Пермского края до 2026 года, созданной Постановлением Законодательного Собрания Пермского края от 16.06.2011 N 2696
Устав Пермского края	перечень показателей результативности деятельности Правительства Пермского края, утвержденный Постановлением Законодательного Собрания

	Пермского края от 22.09.2011 N 2868
Закон Пермского края от 02.04.2010 N 598-ПК "О стратегическом планировании социально-экономического развития Пермского края"	
Федеральный закон от 22.09.2009 N 218-ФЗ	

Представление данного документа в индексе отображено в Приложении 3.

2.5. Описание процесса поиска

Пользователь формирует информационный запрос средствами клиентского приложения, приложение принимает запрос и передает его серверу Solr в соответствующем виде (JSON, XML и др.). Сервер обрабатывает запрос и формирует выборку релевантных документов. Выборка передается клиентскому приложению в формате JSON, приложение преобразует результат и выводит список документов в читаемом формате.

2.5.1. Формирование запросов средствами Solr

В Solr есть собственный пользовательский интерфейс (рис. 11), который позволяет искать документы по сформированным индексам.

Solr предоставляет следующие возможности настройки поискового запроса:

- Поиск по ключевым словам;
- Установка фильтров;
- Выбор полей, отображаемых в результате;
- Фасетный поиск;
- И др.

Таким образом, пользователь имеет возможность гибкой настройки параметров запроса, что позволяет влиять на релевантность результатов

поиска. Однако вывод результатов поиска в Solr осуществляется посредством отображения индексов документа, что затрудняет восприятие результатов.

В связи с этим возникает необходимость разработки пользовательского интерфейса, который позволит формировать результаты поиска в адаптированном для восприятия виде, т.е. вместо вывода содержания индексов документов отображать ссылки на документы, переходя по которым пользователь сможет ознакомиться с его содержанием и списком связей.

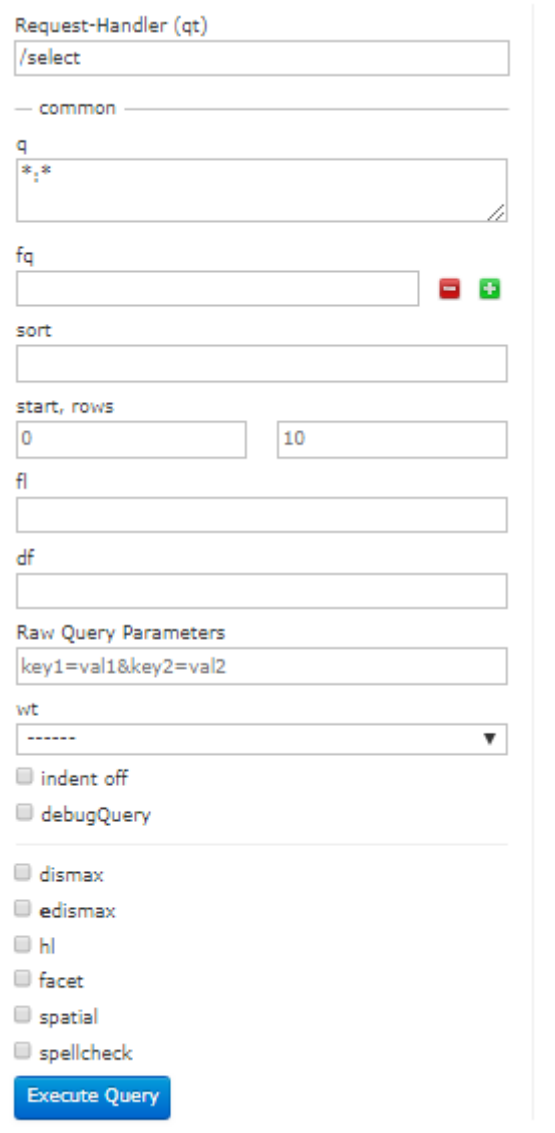

Рис. 11. Встроенный пользовательский интерфейс Solr

2.5.2. Формирование запросов средствами клиентского приложения

Solr поддерживает возможность взаимодействия с приложениями, создаваемыми посредством использования популярных языков программирования таких как Java, C#, Python, PHP, Ruby и другие. В релизе Solr включены библиотеки для взаимодействия с этими средствами разработки. Наиболее подходящим вариантом для создания приложения работы с Solr является разработка web-интерфейса.

Существует несколько способов разработки web-приложения, взаимодействующего с Solr:

- Использование библиотек JavaScript таких как Vue и Axios. Данный способ позволит организовать передачу запросов и ответов от клиента к серверу, путем формирования и передачи HTTPS/JSON-запросов серверу с Solr. Solr возвращает ответ на JSON, который формирует объект JavaScript. Используя свойства полученного объекта можно отображать необходимую информацию о документе.
- Использовать встроенные библиотеки для связи с клиентским приложением созданным средствами определенного языка программирования.
- Использовать CMS Drupal для подключения готового модуля интеграции с Solr. Этот способ позволяет провести экспресс настройку для быстрого начала работы с поисковым сервером, но гибкость настроек при таком использовании отсутствует.

Все вышеописанные способы имеют как преимущества, так и недостатки. Реализация пользовательского интерфейса для взаимодействия с сервисом Solr является весьма трудоёмкой задачей. В данной работе сформирован макет пользовательского интерфейса, который можно использовать в дальнейшем, после настройки всех необходимых модулей подключения.

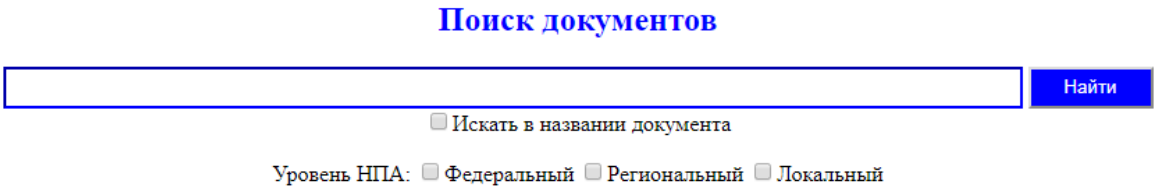

Рис. 12. Стартовая страница интерфейса

В начале работы с пользовательским интерфейсом пользователь вводит запрос на поиск данных в текстовое поле и устанавливает необходимые фильтры (Рис. 12).

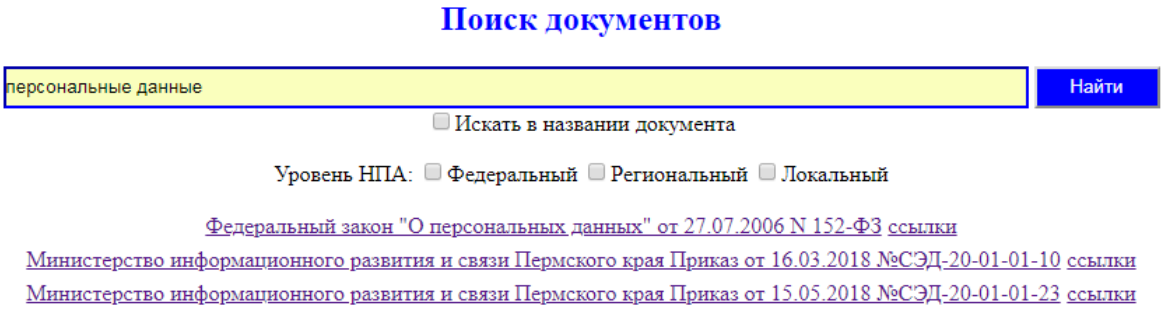

Рис. 13. Результаты обработки запроса пользователя

После отправки запроса серверу, пользователь получает список документов, отвечающих его информационным потребностям (Рис. 13).

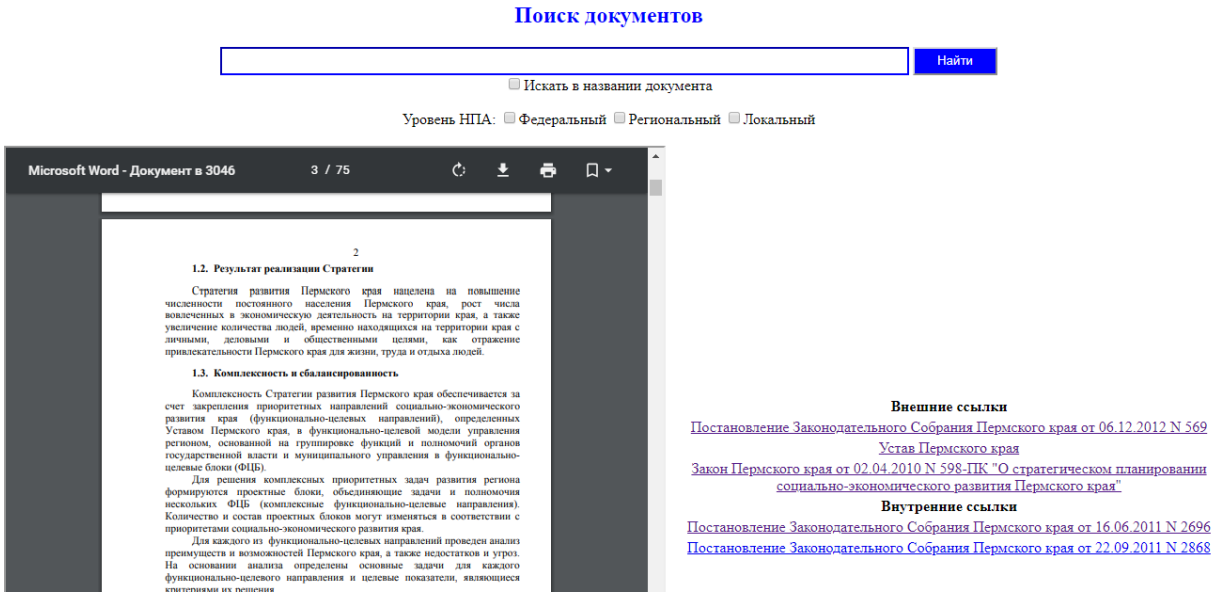

Рис. 14. Отображение содержания документа

Список представляет собой набор гиперссылок, переходя по которым пользователь может просмотреть содержание документа и его ссылки (рис.14).

2.6. Тестирование по запросам

Цель тестирования - проверка работоспособности оболочки:

- Выгрузка входных данных в поисковый сервер;
- Адаптация индексов документов под условия предметной области;
- Обработка поисковых запросов;
- Формирование выборки релевантных документов (осуществление полнотекстового поиска по заданным параметрам).

Тестирование проводилось на выборке нормативных документов, используемых для осуществления деятельности Министерства информационного развития и связи. Документы для выборки отбирались случайным образом. Фрагмент списка НПА:

— Документы федерального уровня:

- Федеральный закон «Об информации, информационных технологиях и о защите информации» от 27.07.2006 N 149-ФЗ;
- Федеральный закон «О персональных данных» от 27.07.2006 N 152-ФЗ;
- Федеральный закон «Об электронной подписи» от 06.04.2011 N 63-ФЗ;
- Федеральный закон «О внесении изменений в Федеральный закон Об информации, информационных технологиях и о защите информации и Федеральный закон Об обеспечении доступа к информации о деятельности государственных органов и органов местного самоуправления» от 07.06.2013 N 112-ФЗ;
- Федеральный закон «Об организации предоставления государственных и муниципальных услуг» от 27.07.2010 N 210-ФЗ;

- Распоряжение Правительства РФ от 28.12.2011 N 2415-р “О государственных и муниципальных услугах, предоставляемых в электронном виде”;
- Распоряжение Правительства РФ от 30.12.2013 N 2602-р “Об утверждении плана мероприятий («дорожной карты») «Развитие отрасли информационных технологий»;
- Постановление Правительства РФ от 08.09.2010 N 697 «О единой системе межведомственного электронного взаимодействия»;
- Постановление Правительства РФ от 21.04.2005 N 241 «О мерах по организации оказания универсальных услуг связи»;
- Распоряжение Правительства РФ от 06.05.2008 N 632-р “О Концепции формирования в Российской Федерации электронного правительства до 2010 года”;
- Распоряжение Правительства РФ от 25.04.2011 N 729-р “Об утверждении перечня услуг, оказываемых государственными и муниципальными учреждениями и другими организациями, в которых размещается государственное задание (заказ) или муниципальное задание (заказ), подлежащих включению в реестры государственных или муниципальных услуг и предоставляемых в электронной форме”;
- Постановление Правительства РФ от 24.10.2011 N 861 «О федеральных государственных информационных системах, обеспечивающих предоставление в электронной форме государственных и муниципальных услуг (осуществление функций)».

— Документы регионального уровня:

- Постановление Законодательного Собрания Пермского края от 1 декабря 2011 года N 3046 “О стратегии социально-экономического развития Пермского края до 2026 года”;
- Постановление Законодательного Собрания Пермского края от 6 декабря 2012 года N 569 “О внесении изменений в Постановление

Законодательного Собрания Пермского края от 01.12.2011 года N 3046 «О стратегии социально-экономического развития Пермского края до 2026 года»”.

— Локальные документы:

- Приказ Министерства информационного развития и связи Пермского края от 16.03.2018 №СЭД-20-01-01-10 “Об осуществлении функций оператора персональных данных в Единой информационной системе управления финансово-хозяйственной деятельностью организаций государственного сектора Пермского края”;
- Приказ Министерства информационного развития и связи Пермского края от 11.04.2018 №СЭД-20-01-01-16 “О внесении изменений в Приказ Министерства информационного развития и связи Пермского края от 23 июня 2016 г. №СЭД-20-01-06-36 “Об утверждении нормативных затрат на обеспечение функций Министерства информационного развития и связи Пермского края”;
- Приказ Министерства информационного развития и связи Пермского края от 15 мая 2018 г. №СЭД-20-01-01-23 “О персональных данных в Министерстве информационного развития и связи Пермского края”.

После формирования массива входных документов, описания схемы и настройки конфигурации можно приступать к автоматическому индексированию средствами Solr. Листинг процесса формирования индексов методом Post представлен в Приложении 4.

Далее автоматические индексы проходят проверку и правку, в конечном итоге в индексе каждого документа содержатся следующие поля:

- Id – идентификатор;
- Author – наименование организации, издавшей документ;
- Date – дата публикации;
- Links – идентификаторы документов, на которые ссылается документ;

- Title – название документа;
- Lvl – законодательный уровень документа (федеральный, региональный, локальный);
- Location – расположение документа (путь к документу на компьютере/в базе/url ссылка);
- Status – действующий/не действующий документ;
- Content – содержание документа;
- TypeDoc – формат исходного документа.

После того, как индексы всех документов сформированы и оптимизированы, можно приступать к поиску. Для обеспечения наибольшей релевантности результатов запросам пользователей помимо реализации полнотекстового поиска предполагается использование фильтров, таких как:

- поиск по названию документа;
- поиск документов, принятых указанной организацией;
- поиск документов указанного законодательного уровня;
- поиск документов с указанным статусом;
- поиск документов определенного формата.

Формирование запросов для тестовой выборки.

1) Поиск по названию документа:

Поисковый запрос “ФЗ 149” в представлении Solr: localhost:8983/solr/file/select?q=title: ФЗ 149

Результат выполнения запроса приведен на рисунке 15.


```
{
  "id":"149-fz-27-07-2006",
  "author":["Государственная Дума РФ"],
  "date":["2006-07-27T00:00:00Z"],
  "title":"Федеральный закон Российской Федерации от 27 июля 2006 г. N 149-ФЗ Об информации, информационных технологи
  "lvl":["Федеральный"],
  "location":["npa/fed/fz/zak_ob_informacii.rtf"],
  "typedoc":["rtf"]},
```


Рис. 15. Результат выполнения запроса “ФЗ 149”

Выборка представлена документом “Федеральный закон «Об информации, информационных технологиях и о защите информации» от 27.07.2006 N 149-ФЗ”.

2) Поиск документов, принятых указанной организацией:

Поисковый запрос “Законодательное собрание Пермского края” в представлении Solr: localhost:8983/solr/file/select?fq=author:”Законодательное собрание Пермского края”&q=*:*

Результат выполнения запроса приведен на рисунке 16.


```
{
  "id":"569-reg-post-06-12-2012",
  "author":["Законодательное Собрание Пермского края"],
  "date":["2012-12-06T00:00:00Z"],
  "title":"Постановление Законодательного Собрания Пермского края от 6 декабря 2012 года N 569 О внесении изменений в
  "lvl":["Региональный"],
  "location":["npa/red/post/22107.pdf"],
  "typedoc":["pdf"]},
{
  "id":"3046-post-pk-01-12-2011",
  "author":["Законодательное Собрание Пермского края "],
  "date":["2011-12-01T00:00:00Z"],
  "title":"ПОСТАНОВЛЕНИЕ от 1 декабря 2011 г. N 3046 О СТРАТЕГИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО РАЗВИТИЯ ПЕРМСКОГО КРАЯ ДО
  "lvl":["региональный"],
  "location":["npa\regpost/Postanovlenije_ZS_PK_ot_1.12.2011_N3046.doc"],
  "typedoc":["doc"]},
```


Рис. 16. Поиск документов принятых Законодательным собранием Пермского края

Выборка представлена документами “Постановление Законодательного Собрания Пермского края от 6 декабря 2012 года N 569 “О внесении изменений в Постановление Законодательного Собрания Пермского края от 01.12.2011 года N 3046 «О стратегии социально-экономического развития Пермского края до 2026 года»” и “ Постановление Законодательного Собрания Пермского края от 1 декабря 2011 года N 3046 “О стратегии социально-экономического развития Пермского края до 2026 года”.

3) Поиск документов указанного законодательного уровня:

Поисковый запрос “Локальный” в представлении Solr: http://localhost:8983/solr/file/select?fq=lvl:”Локальный”&q=*:*

Результат выполнения запроса приведен на рисунке 17.

```
{
  "id":"СЭД-20-01-01-10-loc-prik-16-03-2018",
  "author":["Министерство инфокоммуникационного развития и связи Пермского края"],
  "date":["2018-03-16T00:00:00Z"],
  "title":"Приказ Министерства информационного развития и связи Пермского края от 16.03.2018 №СЭД-20-01-01-10 Об осущ
  "lvl":["Локальный"],
  "location":["npa/loc/prikaz/Пр СЭД-20-01-01-10 от 16.03.2018 Об осуществлении функций оператора.pdf"],
  "typedoc":["pdf"]},
{
  "id":"СЭД-20-01-01-16-loc-prik-11-04-2018",
  "author":["Министерство инфокоммуникационного развития и связи Пермского края"],
  "date":["2018-04-11T00:00:00Z"],
  "title":"Приказ Министерства информационного развития и связи Пермского края от 11.04.2018 №СЭД-20-01-01-16 О внесе
  "lvl":["Локальный"],
  "location":["npa/loc/prikaz/Пр СЭД-20-01-01-16 от 11.04.2018 О внесении изменений в приказ.pdf"],
  "typedoc":["pdf"]},
{
  "id":"СЭД-20-01-01-23-loc-prik-15-05-2018",
  "author":["Министерство инфокоммуникационного развития и связи Пермского края"],
  "date":["2018-04-11T00:00:00Z"],
  "title":"Приказ Министерства информационного развития и связи Пермского края от 15 мая 2018 г. №СЭД-20-01-01-23 О п
  "lvl":["Локальный"],
  "location":["npa/loc/prikaz/Пр СЭД-20-01-01-23 от 15.05.2018 О персональных данных.pdf"],
  "typedoc":["pdf"]}]
```


Рис. 17. Поиск локальных НПА

Выборка представлена документами “Приказ Министерства информационного развития и связи Пермского края от 16.03.2018 №СЭД-20-01-01-10 “Об осуществлении функций оператора персональных данных в Единой информационной системе управления финансово-хозяйственной деятельностью организаций государственного сектора Пермского края”, “Приказ Министерства информационного развития и связи Пермского края от 11.04.2018 №СЭД-20-01-01-16 “О внесении изменений в Приказ Министерства информационного развития и связи Пермского края от 23 июня 2016 г. №СЭД-20-01-06-36 “Об утверждении нормативных затрат на обеспечение функций Министерства информационного развития и связи Пермского края” и “ Приказ Министерства информационного развития и связи Пермского края от 15 мая 2018 г. №СЭД-20-01-01-23 “О персональных данных в Министерстве информационного развития и связи Пермского края”.

4) Комбинированные настройки

Поисковый запрос “персональные данные” с фильтром “уровень документа: локальный” в представлении Solr: localhost:8983/solr/file/select?fq=lvl:”Локальный”&q=content:персональные данные*

Результат выполнения запроса приведен на рисунке 18.


```
{
  "id":"СЭД-20-01-01-23-loc-prik-15-05-2018",
  "author":["Министерство инфокоммуникационного развития и связи Пермского края"],
  "date":["2018-04-11T00:00:00Z"],
  "links_out":["-",
    "-",
    "-",
    "-"],
  "links_input":["-",
    "-"],
  "title":"Приказ Министерства информационного развития и связи Пермского края от 15 мая 2018 г. №СЭД-20-01-01-23 О п
  "lvl":["Локальный"],
  "location":["npa/loc/prikaz/Пр СЭД-20-01-01-23 от 15.05.2018 О персональных данных.pdf"],
  "status":["Действующий"],
  "typedoc":["pdf"],
  "content":"Персональные данные правила обработки персональных данных в Министерстве инфокоммуникационного развития
  "language":"ru",
  "_version_":1604412385807826944}]
```


Рис. 18. Поиск локальных НПА по запросу “персональные данные”

Выборка представлена документом “ Приказ Министерства информационного развития и связи Пермского края от 15 мая 2018 г. №СЭД-20-01-01-23 “О персональных данных в Министерстве информационного развития и связи Пермского края ”.

Испытания, проведенные на тестовой выборке документов, показали, что алгоритм описывающий работоспособность системы возможно реализовать на каждом этапе. Результат проведенного поиска соответствует ожидаемому.

Глава 3. Анализ организационного финансового обоснования проекта

В процессе внедрения информационных технологий возникает необходимость оценки экономической эффективности информационного продукта. Существует множество различных методик по оценке эффективности внедрения IT-продукта (рис. 19).

Рис. 19. Классификация методов оценки экономической эффективности ИТ-проектов [5]

Наиболее эффективным является метод TCO (совокупной стоимости владения). Данный метод позволяет оценить затраты связанные с приобретением, внедрением и использованием информационной системы. При этом затраты подразделяются на две группы: фиксированные затраты и текущие.

К фиксированным относятся следующие затраты:

- стоимость разработки и внедрения проекта;
- привлечение внешних консультантов;
- первоначальные закупки основного ПО;

- первоначальные закупки дополнительного ПО;
- первоначальные закупки аппаратного обеспечения.

Текущие затраты состоят из трех статей:

- стоимость обновления и модернизации системы;
- затраты на управление системой в целом;
- затраты, вызванные активностью пользователей ИС.

В общем случае *TCO* оценивается по формуле:

$$TCO = K + n \times C[\text{руб.}], \quad (1)$$

где C – эксплуатационные затраты на ИС;
K – капитальные (единовременные) затраты на ИС;
n – количество планируемых лет эксплуатации ИС.

Так как для разработки используется личный компьютер разработчика, а для реализации уже существующие технические средства, затраты на приобретение техники в данной разработке отсутствуют. Для разработки используется бесплатное программное обеспечение: Solr, Denwer и Ramus Educantional. Таким образом, на этапе реализации учитываются только затраты на выплату заработной платы.

3.1. Расчет единовременных затрат на разработку оболочки поисково-аналитической системы.

Перед проектированием и разработкой системы необходимо провести анализ предметной области, выявить особенности входной документации, а также определить требования к проектируемой системе. По данным сайта perm.trud.com средняя зарплата аналитика в Пермском крае составляет в среднем 25 000 рублей в месяц (25000/(22*8)=142 руб/час).

На стадии проектировании и разработки необходимо: проанализировать существующие программные средства, направленные на решение поставленных задач; изучить техническую документацию для работы с выбранным программным продуктом; реализовать процесс

автоматизированного построения полнотекстовых индексов документов; разработать пользовательский интерфейс. Средняя зарплата разработчика ИС в Пермском крае составляет 45 000 рублей в месяц (45000/(22*8)=255 руб/час)

Расчет единовременных затрат сведен в таблицу 5.

Таблица 5. Единовременные затраты на разработку оболочки поисково-аналитической системы

Процесс	Трудоемкость (час)	Ставка исполн., час	Стоимость
Анализ предметной области	16	142	2272
Анализ существующих программных продуктов, поддерживающих реализацию функции полнотекстового поиска	5	255	1275
Анализ нормативно-правовой документации	32	142	4544
Проектирование требований к системе	2	142	284
Анализ технической документации используемого ПО	15	255	3825
Разработка полнотекстовых индексов для нормативной базы организации	40	255	10200
Разработка пользовательского интерфейса	24	255	6120

Итого	134		28520

3.2. Расчет текущих затрат на эксплуатацию оболочки поисково-аналитической системы.

Внедрение, эксплуатация и сопровождение осуществляется администратором системы. Зарплата за месяц около 25000 руб, 25000/(22*8)=142 руб/час. На этапе внедрения необходимо установить и настроить систему, затем постепенно заполнить базу системы необходимыми документами, на это потребуется примерно 5 недель (25*8=200 часов). На этапе эксплуатации необходимо будет раз в месяц обновлять базу документов, обновления займут по 2 часа в месяц (за год 12*2=24 часа). Стоимость сопровождения составляет 10% от затрат на этапах исследования и реализации: (28520*10)/100=2852 руб/мес, за год 34224 руб/год.

Затраты на внедрение и эксплуатацию системы сведены в таблице 6.

Таблица 6. Текущие затраты на использование оболочки поисково-аналитической системы за год

Процесс	Трудоемкость (час)	Ставка исполн., час	Стоимость
Установка и настройка ПО	2	142	284
Подготовка входных данных	8	142	1136
Обработка индексов документов	190	142	26980
Администрирование ИС	24	142	3408
Техническое обслуживание	-	-	34224
Итого	224		66032

Выводы по третьей главе

Таким образом, совокупная стоимость разрабатываемого программного продукта составляет 94552 рублей при сроке эксплуатации равном одному году. Согласно сайту zakupki.gov.ru стоимость развития существующих информационных систем, направленных на обеспечение процесса поиска и анализа нормативной базы документов, в среднем составляет 1000000 рублей, сопровождение и обслуживание поисковых систем 500000 рублей.

Разрабатываемая система имеет ряд преимуществ относительно аналогов, среди явных преимуществ можно выделить следующие:

- Программный продукт реализован с использованием свободно распространяемого ПО с открытым исходным кодом;
- Возможна индексация документов из различных источников, таких как сайты в сети Интернет, базы данных, облачные хранилища, почтовые серверы, локальные директории;
- Поддержка различных форматов документов (XML, CSV, JSON, RTF, PDF, TXT, форматы офисных документов, таких как текстовые документы, электронные таблицы, презентации и пр.);
- Кроссплатформенность: и серверная и клиентская часть поддерживается на всех широко распространенных ОС;
- Гибкая настройка, что позволяет адаптировать программный продукт практически под любые нужды организации и потребности пользователей.

Ввиду сложности теоретической оценки эффекта от внедрения разрабатываемого программного продукта, на данном этапе не представляется возможным обосновать эффективность использования оболочки поисково-аналитической системы. Как и любой другой IT-продукт, данная разработка должна пройти этап эксплуатации для выявления

объективных факторов влияющих на оценку эффективности применения программного средства.

Заключение

В результате выполнения выпускной квалификационной работы было проведено проектирование и разработка оболочки поисково-аналитической системы поддержки правовой деятельности органов государственной власти на примере Министерства информационного развития и связи Пермского края. Оболочка нацелена на выполнение задач полнотекстового поиска и анализа связей между документами нормативной базы организации. Основной целью работы было создание системы индексирования слабо формализованной информации.

В результате анализа предметной области и источников литературы была сформирована выборка нормативной документации, сформулированы требования к результатам (реализация полнотекстового поиска по заданным параметрам с указанием связей между документами), определившие структуру ПО (поисковый сервер - Solr, модуль обработки запросов, модуль анализа и оптимизации индексов документов) и алгоритм его работы.

Для тестирования разработанной оболочки была сформирована выборка нормативной документации, на примере которой был апробирован алгоритм работы оболочки. Результаты тестирования показали, что разработанная архитектура ПО отвечает требованиям полнотекстового поиска и анализа связей в массиве слабо формализованных дынных.

После соответствующих доработок оболочка поисково-аналитической системы может применяться в решении разнообразного класса теоретических и практических задач.

Список источников

1. Комплексный анализ и оценка информационной открытости сайтов органов государственного управления Пермского края. Сборник научных статей / Под ред. доктора экономических наук, профессора Н. Л. Казариновой. – Пермь: АНО ДПО «ПИМУиИ». – 84 с
2. Автоматизированные информационные системы, базы и банки данных. Вводный курс: Учебное пособие. — М.: Гелиос АРВ, 2002. — 368 с.
3. Официальный сайт Министерства информационного развития и связи Пермского края. URL: http://mirs.permkrai.ru/ (дата обращения 24.06.2018)
4. Sachin Handiekar, Anshul Johri (2015) Apache Solr for Indexing Data. Published by Packt Publishing Ltd. Livery Place 35 Livery Street Birmingham B3 2PB, UK.
5. Учебник 4CIO. Версия 2.0. Под редакцией С. Кирюшина. Москва. 4CIO. 2013 – 700 с.
6. Сайт вакансий города Перми. URL: http://perm.trud.com (дата обращения 24.06.2018)
7. Официальный сайт единой информационной системы в сфере закупок http://zakupki.gov.ru/epz/main/public/home.html (дата обращения 24.06.2018)
8. ГОСТ 7.66-92 “Индексирование документов”
9. Основы программной инженерии (по SWEBOK) - перевод SWEBOK 2004 с замечаниями и комментариями, подготовленный Сергеем Орликом. URL: http://swebok.sorlik.ru
10. Apache Solr Wiki. URL: https://wiki.apache.org/solr/ (дата обращения 18.04.2018)

Приложение 1.

Приложение 2.
Представление документа в индексе до обработки

{

"id":"/home/daria/Документы/npa/reg/post/Postanovlenije_ZS_PK_ot_1.12.2011_N3046.doc",

"attr_cp_revision":["1"],

"attr_date":["2014-08-04T04:59:00Z"],

"attr_stream_content_type":["application/msword"],

"attr_meta_word_count":["28406"],

"attr_dc_creator":["Пользователь"],

"attr_word_count":["28406"],

"attr_dcterms_created":["2014-08-04T04:57:00Z"],

"attr_dcterms_modified":["2014-08-04T04:59:00Z"],

"attr_last_modified":["2014-08-04T04:59:00Z"],

"attr_last_save_date":["2014-08-04T04:59:00Z"],

"attr_meta_character_count":["161919"],

"attr_template":["Normal.dotm"],

"attr_meta_save_date":["2014-08-04T04:59:00Z"],

"attr_application_name":["Microsoft Office Word"],

"attr_modified":["2014-08-04T04:59:00Z"],

"attr_edit_time":["1200000000"],

"content_type":"application/msword",

"attr_stream_size":["686592"],

"attr_x_parsed_by":["org.apache.tika.parser.DefaultParser",

"org.apache.tika.parser.microsoft.OfficeParser"],

"attr_creator":["Пользователь"],

"attr_meta_author":["Пользователь"],

"attr_extended_properties_application":["Microsoft Office Word"],

"attr_meta_creation_date":["2014-08-04T04:57:00Z"],

"attr_meta_last_author":["Пользователь"],

"attr_creation_date":["2014-08-04T04:57:00Z"],

"attr_xmptpg_npages":["83"],

"attr_resourcename":["/home/daria/Документы/npa/reg/post/Postanovlenije_ZS_PK_ot_1.12.2011_N3046.doc"],

"attr_last_author":["Пользователь"],

"attr_character_count":["161919"],

"attr_page_count":["83"],

"attr_revision_number":["1"],

"attr_extended_properties_template":["Normal.dotm"],

"attr_author":["Пользователь"],

"attr_meta_page_count":["83"],

"language":"ru",

"content_type_type_s":"application",

"content_type_subtype_s":"msword",

"url_ss":["http://www.infomine.com/ChartsAndData"],

"_version_":1603146718807851008,

"content": " обычный ЗАКОНОДАТЕЛЬНОЕ СОБРАНИЕ ПЕРМСКОГО КРАЯ \n обычный ПОСТАНОВЛЕНИЕ\n \n обычный от 1 декабря 2011 г. N 3046\n \n обычный О СТРАТЕГИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО РАЗВИТИЯ

ПЕРМСКОГО КРАЯ\n \n обычный ДО 2026 ГОДА\n \n обычный (в ред. Постановления Законодательного Собрания\n \n обычный Пермского края от 06.12.2012 N 569)\n \n обычный Законодательное Собрание Пермского края постановляет:\n \n обычный 1. Утверд Стратег соц-эконом развит Перм кр до 2026 года Пост Законодательного Собрания Перм кр от 16.06.2011 N 2696.\n \n обычный 2. Контроль исполнение пост комитет Законодательн Собран Перм кр гос политик мест самоуправлен эконом полит природопользован.\n \n обычный 3. Пост сил принят.\n \n обычный Председатель\n \n обычный Законодательн Собрани\n \n обычный Н.А.ДЕВЯТКИН\n \n обычный Приложение\n \n обычный Постановлен\n \n обычный Законодательн Собрани\n \n обычный Перм кр\n \n обычный 01.12.2011 N 3046\n \n обычный СТРАТЕГ\n \n обычный СОЦ ЭКОНОМ РАЗВИТИЯ ПЕРМ КР\n \n обычный 2026 \n \n обычный Перм кр 06.12.2012 N 569\n \n обычный 1. ЦЕЛЬ СТРАТЕГ РАЗВИТ ПЕРМ КР\n \n обычный 1.1. Стратег\n \n обычный Стратег развит комплекс и сбалансирован развит Пермс кр показателей.\n \n обычный Цель формирован Стратег обеспечен комплекс сбалансирован развит Перм кр.\n \n обычный Под набором общественно признанных показателей для целей определения Стратегии принимается перечень показателей результативности деятельности Правительства Пермского края, утвержденный Постановлением Законодательного Собрания Пермского края от 22.09.2011 N 2868.\n \n. Оценк изменен набор показател.\n \ 1.2. Результат реализации Стратегии\n \n обычный \n обычный 1.3. Комплексность и сбалансированность\ \n обычный 1.4. Период реализации Стратегии и ее корректировка\n \ ОСНОВНЫЕ НАПРАВЛЕНИЯ И ЗАДАЧИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО\n \n обычный РАЗВИТИЯ ПЕРМСКОГО КРАЯ\n Социальная политика\";\n \n обычный \"Общественная безопасность\";\n \n обычный \"Экономическая политика\";\n \n обычный \"Природопользование и инфраструктура\";\n \n обычный \"Управление земельными ресурсами и имуществом\";\n \n обычный \"Территориальное развитие\".\n \n обычный 3. СТРАТЕГИЧЕСКИЕ ЦЕЛИ, ЗАДАЧИ И ПОКАЗАТЕЛИ\n \n обычный ЦЕЛЕВЫЕ ПОКАЗАТЕЛИ\n \n обычный РЕЗУЛЬТАТИВНОСТИ ДЕЯТЕЛЬНОСТИ ПРАВИТЕЛЬСТВА ПЕРМСКОГО КРАЯ\n \ \n обычный Перечень\n \n обычный ключевых проектных блоков и проектов Пермского края\n \n обычный Исключен с 6 декабря 2012 года. - Постановление Законодательного Собрания Пермского края от 06.12.2012 N 569.\n \n обычный Приложение 3\n \n обычный к Стратегии\n \n обычный

АНАЛИЗ\n \n обычный СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО РАЗВИТИЯ ПЕРМСКОГО КРАЯ\n \n"

}

Приложение 3.
Представление документа в индексе после обработки

{

"id": "3046-post-pk-01-12-2011",

"author": "Законодательное Собрание Пермского края ",

"date": "2011-12-01",

"links_out": [

"569-post-pk-06-12-2012",

"us-pk",

"598-zk-pk-02-04-2010",

"218-fz-22-09-2009"

],

"links_input": [

"2696-post-pk-22-09-2011",

"2868-post-pk-22-09-2011"

],

"title": "ПОСТАНОВЛЕНИЕ от 1 декабря 2011 г. N 3046 О СТРАТЕГИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО РАЗВИТИЯ ПЕРМСКОГО КРАЯ ДО 2026 ГОДА",

"lvl": "региональный",

"location": "npa\reg\post/Postanovlenije_ZS_PK_ot_1.12.2011_N3046.doc",

"content": " обычный ЗАКОНОДАТЕЛЬНОЕ СОБРАНИЕ ПЕРМСКОГО КРАЯ \n обычный ПОСТАНОВЛЕНИЕ\n \n обычный от 1 декабря 2011 г. N 3046\n \n обычный О СТРАТЕГИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО РАЗВИТИЯ ПЕРМСКОГО КРАЯ\n \n обычный ДО 2026 ГОДА\n \n обычный (в ред. Постановления Законодательного Собрания\n \n обычный Пермского края от 06.12.2012 N 569)\n \n обычный Законодательное Собрание Пермского края постановляет:\n \n обычный 1. Утверд Стратег соц-эконом развит Перм кр до 2026 года Пост Законодательного Собрания Перм кр от 16.06.2011 N 2696.\n \n обычный 2. Контроль исполнение пост комитет Законодательн Собран Перм кр гос политик мест самоуправлен эконом полит природопользован.\n \n обычный 3. Пост сил принят.\n \n обычный Председатель\n \n обычный Законодательн Собрани\n \n обычный Н.А.ДЕВЯТКИН\n \n обычный Приложение\n \n обычный Постановлен\n \n обычный Законодательн Собрани\n \n обычный Перм кр\n \n обычный 01.12.2011 N 3046\n \n обычный СТРАТЕГ\n \n обычный СОЦ ЭКОНОМ РАЗВИТИЯ ПЕРМ КР\n \n обычный 2026 \n \n обычный Перм кр 06.12.2012 N 569\n \n обычный 1. ЦЕЛЬ СТРАТЕГ РАЗВИТ ПЕРМ КР\n \n обычный 1.1. Стратег\n \n обычный Стратег развит комплекс и сбалансирован развит Пермс кр показателей.\n \n обычный Цель формирован Стратег обеспечен комплекс сбалансирован развит Перм кр.\n \n обычный Под набором общественно признанных показателей для целей определения Стратегии принимается перечень показателей результативности деятельности Правительства Пермского края, утвержденный Постановлением Законодательного Собрания Пермского края от 22.09.2011 N 2868.\n \n. Оценк изменен набор показател.\n \ 1.2. Результат реализации Стратегии\n \n обычный \n обычный 1.3. Комплексность и сбалансированность\ \n обычный 1.4. Период реализации Стратегии и ее корректировка\n \ ОСНОВНЫЕ НАПРАВЛЕНИЯ И ЗАДАЧИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО\n \n обычный РАЗВИТИЯ ПЕРМСКОГО КРАЯ\n Социальная политика\";\n \n обычный \"Общественная безопасность\";\n \n обычный \"Экономическая политика\";\n \n обычный \"Природопользование и инфраструктура\";\n \n обычный \"Управление земельными ресурсами и имуществом\";\n \n обычный \"Территориальное развитие\".\n \n обычный 3. СТРАТЕГИЧЕСКИЕ ЦЕЛИ, ЗАДАЧИ И ПОКАЗАТЕЛИ\n \n обычный ЦЕЛЕВЫЕ ПОКАЗАТЕЛИ\n \n обычный РЕЗУЛЬТАТИВНОСТИ ДЕЯТЕЛЬНОСТИ ПРАВИТЕЛЬСТВА ПЕРМСКОГО КРАЯ\n \ \n обычный Перечень\n \n обычный ключевых проектных блоков и

проектов Пермского края\n \n обычный Исключен с 6 декабря 2012 года. - Постановление Законодательного Собрания Пермского края от 06.12.2012 N 569.\n \n обычный Приложение 3\n \n обычный к Стратегии\n \n обычный АНАЛИЗ\n \n обычный СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО РАЗВИТИЯ ПЕРМСКОГО КРАЯ\n \n"

}

Приложение 4.
Листинг процесса индексации документов методом Post

daria@daria-VirtualBox ~/Загрузки/solr-7.3.1 $ bin/post -c npa /home/daria/Документы/npa

/usr/lib/jvm/java-8-oracle/bin/java -classpath /home/daria/Загрузки/solr-7.3.1/dist/solr-core-7.3.1.jar -Dauto=yes -Dc=npa -Ddata=files -Drecursive=yes org.apache.solr.util.SimplePostTool /home/daria/Документы/npa

SimplePostTool version 5.0.0

Posting files to [base] url http://localhost:8983/solr/npa/update...

Entering auto mode. File endings considered are xml,json,jsonl,csv,pdf,doc,docx,ppt,pptx,xls,xlsx,odt,odp,ods,ott,otp,ots,rtf,htm,html,txt,log

Entering recursive mode, max depth=999, delay=0s

Indexing directory /home/daria/Документы/npa (2 files, depth=0)

POSTing file open_data-4.xml (application/xml) to [base]

POSTing file document.xml (application/xml) to [base]

Indexing directory /home/daria/Документы/npa/reg (0 files, depth=1)

Indexing directory /home/daria/Документы/npa/reg/ukaz (0 files, depth=2)

Indexing directory /home/daria/Документы/npa/reg/post (3 files, depth=2)

POSTing file Postanovlenije_ZS_PK_ot_1.12.2011_N3046.doc (application/msword) to [base]/extract

POSTing file 22107.pdf (application/pdf) to [base]/extract

POSTing file 911543881.pdf (application/pdf) to [base]/extract

Indexing directory /home/daria/Документы/npa/reg/prikaz (0 files, depth=2)

Indexing directory /home/daria/Документы/npa/fed (0 files, depth=1)

Indexing directory /home/daria/Документы/npa/fed/ukaz (0 files, depth=2)

Indexing directory /home/daria/Документы/npa/fed/post (7 files, depth=2)

POSTing file post-anovlenie_pr_va_rf_241_ot_21_04_2005_o_merah_po_organizacii_okazaniya_universal_nyh_uslug_svyazi.pdf (application/pdf) to [base]/extract

POSTing file Post_697_1.pdf (application/pdf) to [base]/extract

POSTing file Rasp_Prav_RF_N_729-r.pdf (application/pdf) to [base]/extract

POSTing file Постановление Правительства РФ от 24.10.2011 N 861.rtf (text/rtf) to [base]/extract

POSTing file 5izmeneniya2415-r.rtf (text/rtf) to [base]/extract

POSTing file rasp_6-05-2008_N632-r_1.pdf (application/pdf) to [base]/extract

POSTing file 41d4b29db7c74fb9ad46.pdf (application/pdf) to [base]/extract

Indexing directory /home/daria/Документы/npa/fed/prikaz (0 files, depth=2)

Indexing directory /home/daria/Документы/npa/fed/fz (9 files, depth=2)

POSTing file 152.pdf (application/pdf) to [base]/extract

POSTing file zak_ob_informacii.doc (application/msword) to [base]/extract

POSTing file open_data-4.pdf (application/pdf) to [base]/extract

POSTing file fz-63.pdf (application/pdf) to [base]/extract

POSTing file zak_ob_informacii.xml (application/xml) to [base]

POSTing file ~$k_ob_informacii.xml (application/xml) to [base]

SimplePostTool: WARNING: Solr returned an error #400 (Bad Request) for url: http://localhost:8983/solr/npa/update

SimplePostTool: WARNING: Response: <?xml version="1.0" encoding="UTF-8"?>

<response>

<lst name="responseHeader">

<int name="status">400</int>

<int name="QTime">366</int>

</lst>

<lst name="error">

<lst name="metadata">

<str name="error-class">org.apache.solr.common.SolrException</str>

<str name="root-error-class">java.io.CharConversionException</str>

</lst>

<str name="msg">Invalid UTF-8 start byte 0x98 (at char #135, byte #-1)</str>

<int name="code">400</int>

</lst>

</response>

SimplePostTool: WARNING: IOException while reading response: java.io.IOException: Server returned HTTP response code: 400 for URL: http://localhost:8983/solr/npa/update

POSTing file zak_ob_informacii.rtf (text/rtf) to [base]/extract

POSTing file document.pdf (application/pdf) to [base]/extract

POSTing file pub_279677.pdf (application/pdf) to [base]/extract

Indexing directory /home/daria/Документы/npa/loc (0 files, depth=1)

Indexing directory /home/daria/Документы/npa/loc/prikaz (8 files, depth=2)

POSTing file Пр СЭД-20-01-03-15 от 11.04.2018 Об утверждении нормативных затрат и значений норм, используемых для определения нормативных затрат.pdf (application/pdf) to [base]/extract

POSTing file Пр СЭД-20-01-01-16 от 11.04.2018 О внесении изменений в приказ.pdf (application/pdf) to [base]/extract

POSTing file Пр_СЭД_20_01_02_13_от_16_02_2018.pdf (application/pdf) to [base]/extract

POSTing file Пр СЭД-20-01-03-16 от 11 04 2018 Об утверждении порядков определения нормативных затрат на выполнение работ определения и утверждения норм выраже.pdf (application/pdf) to [base]/extract

POSTing file Приказ от 25.04.2018 г. № СЭД-20-01-02-25 О внесении изменений в Приказ Министерства.pdf (application/pdf) to [base]/extract

POSTing file Пр СЭД-20-01-01-23 от 15.05.2018 О персональных данных.pdf (application/pdf) to [base]/extract

POSTing file Пр СЭД-20-01-01-17 от 12.04.2018 О внесении изменений в План ИТ.PDF (application/pdf) to [base]/extract

POSTing file Пр СЭД-20-01-01-10 от 16.03.2018 Об осуществлении функций оператора .pdf (application/pdf) to [base]/extract

29 files indexed.

COMMITting Solr index changes to http://localhost:8983/solr/npa/update...

Time spent: 0:01:09.710