**Введение**

За последние годы отмечается значительное увеличение объемов информации, представленной, в том числе, в форме текстовых документов. Этим обуславливается потребность в автоматизации различных задач, связанных с обработкой и анализом текстовых данных на естественном языке. Сказанное позволяет сделать вывод об актуальности тематики настоящей выпускной квалификационной работы, связанной с обработкой текстовых документов.

В современной науке выделились следующие направления, связанные с обработкой полнотекстовых документов: машинный перевод, информационной поиск, реферирование текста, классификация текстов, извлечение знаний. В настоящее время направление машинного перевода активно развивается силами компаний-разработчиков систем машинного перевода (PROMT, trados).

Объектом исследования курсовой работы является процесс обработки полнотекстовых документов. Предмет исследования составляют методы информационного поиска и классификации.

Целью курсовой работы является реализация информационно-аналитической системы обработки полнотекстовых документов, реализующей методы информационного поиска и кластеризации. Достижение поставленной цели предполагает решение следующих задач:

1. Систематизировать теоретические сведения о содержании процесса обработки полнотекстового документа;

2. Рассмотреть алгоритмы и инструменты, предназначенные для решения отдельных задач по обработке полнотекстовых документов;

3. Разработать в нотации языка UML проектные решения для программной реализации отдельных алгоритмов обработки текстов;

4. Программно реализовать подготовленные проектные решения.

В ходе выполнения курсовой работы применялись следующие теоретические и эмпирические методы проведения исследования: аппарат математической статистики, алгоритмы стемминга, алгоритмы классификации текстов, методы информационного поиска, возможности языка Python.

Курсовая работа включает в себя: введение, первая глава, вторая глава, заключение и список использованной литературы.

Во введении обоснована актуальность темы, определены объект, предмет исследования, сформулированы цель и задачи курсовой работы, определена структура работы.

В первой главе работы рассматриваются перспективы создания информационно-аналитической системы для обработки полнотекстовой информации, приводятся методы анализа текста, которые должны быть реализованы данной системой.

Во второй главе представлены результаты проектирования системы анализа текста. В частности, приведены ключевые диаграммы, построенные в рамках нотации UML. Описана программная реализация автоматического распределения статей по экспертам.

В заключении сделаны основные выводы по результатам работы. Список литературы содержит n источников, используемых при написании работы.

**1 Анализ предметной области и формирование требований к информационной системе**

* 1. **Актуальность выбранной тематики**

Развитие методов записи и хранения данных привело к бурному росту объемов собираемой и анализируемой информации. Объемы данных настолько внушительны, что человеку просто не по силам проанализировать их самостоятельно, и хотя необходимость проведения такого анализа вполне очевидна, проведем оценку актуальности выбранной тематики. Для этого прибегнем к рассмотрению двух основополагающих для данной предметной области терминов: структурированные и неструктурированные данные.

Неструктурированные данные – данные, не имеющие заранее определенной структуры или не организованные в определенном порядке. Примером неструктурированных данных является текстовая информация. Отсутствие структурированности данных приводит к трудностям анализа, особенно в случае использования традиционных программ, предназначенных для работы со структурированными данными (аннотированными или хранящимися в базах). Помимо текстовых данных, примерами неструктурированных данных являются также электронные сообщения, аудиофайлы, цифровые изображения и видеоклипы. Хотя во всех таких файлах есть некоторая структура, например, в электронных сообщениях есть адрес, тема, текст письма и т. д., которые обычно хранятся в форме, не позволяющей осуществлять простую и логичную классификацию, в отличие от данных, полученных посредством ввода текстовой информации в электронные формы (стандартный способ ручного ввода структурированных данных), в результате вычислений или каких-либо других компьютерных транзакций, в процессе которых автоматически создаются наборы структурированной информации. Пример программ: Пакет программ Microsoft Office, LaTeX.

Структурированные данные – это данные, которые были упорядочены и организованны с целью обеспечения возможности применения к ним алгоритмов обработки. В этом случае подразумевается, что данные упорядочены в виде вертикальных столбцов, именуемых полями, с горизонтальными строками, называемыми записями. Причем все записи содержат один и тот же набор полей, а все поля – один и тот же набор записей.

Обработка неструктурированной информации является актуальной задачей, как минимум, по трем причинам:

1. Деятельность по обработке неструктурированной информация уже привела к значительному продвижению в области ее структуризации, результатами такой деятельности является XML и прочие средства теговой разметки, которые значительно упрощают процесс поиска, классификации, сортировки и создания отчетов для информации, хранящейся в файлах, а не в структурированных базах данных.

2. Проблемы обработки информации, связанные с доступом к файлам, их сохранением, сегодня становятся все менее острыми благодаря непрекращающейся уже более десяти лет работе по отладке операционных систем и открытых стандартов в области извлечения и хранения данных.

3. Системы обработки неструктурированной информации пополняются все новыми функциями, облегчающими использование данной информации для бизнес-целей, например, для прогнозирования и принятия решений.

Для проведения анализа неструктурированных данных применяется Data Mining – процесс выявления скрытых закономерностей, обнаружения в сырых данных (RAW data) ранее неизвестных знаний, простых для интерпретации и практически полезных в различных областях деятельности. Методы Data Mining основываются на базе различных научных дисциплин: статистки, теории баз данных, искусственного интеллекта, алгоритмизации, визуализации и других наук. Алгоритмы, используемые в Data Mining, требуют большого количества вычислений. Раньше это являлось сдерживающим фактором широкого практического применения Data Mining, однако рост производительности современных процессоров снял остроту этой проблемы. Поэтому теперь за небольшой промежуток времени можно провести качественный анализ сотен тысяч и миллионов записей. Одним из разделов Data Mining является Text Mining. Рассмотрим данное направление более подробно, поскольку данная справка пригодится нам далее.

Text Mining – это набор технологий и методов, предназначенных для извлечения информации из текстовых данных. Основной целью технологии Text Mining является обработка больших объемов исходных данных и автоматизация процесса извлечения необходимой информации из обработанного массива данных.

Рассмотрим преимущества и недостатки, а также проведем сравнительный анализ трех подходов к автоматизации процесса получения информации:

• Подход, основанный на применении информационно-поисковых систем (ИПС);

• Подход, основанный на автоматизации извлечения информации;

• Подход, основанный на автоматизации аналитических структур.

Основные преимущества подхода, основанного на применении в различного рода программных продуктах инструмента ИПС, вполне очевидны: распространенность и общедоступность поисковых технологий, большинство решений по созданию поискового инструмента в системе уже реализованы в виде свободно распространяемых и портируемых библиотек. С точки зрения конечного пользователя продукта ИПС – также наиболее доступное и понятное, так называемое one-clickрешение, когда пользователь набирает поисковый запрос, ИПС выдает результаты, релевантные данному запросу, и пользователь вручную обрабатывает всю найденную информацию. Поскольку инструменты поиска развиваются уже давно и достигли высокой стадии зрелости, они вполне успешно отвечают на вопрос, где находится информация. А пользователи успели настолько привыкнуть к интерфейсу ИПС, что нет необходимости проводить какое-то специальное обучение. Однако, если речь идет об обработке больших массивов данных, применение одних только поисковых систем, с точки зрения конечного пользователя, становится малоэффективным, так как требует значительных человеческих ресурсов на этапах поиска информации и ее анализа.

Подход, основанный на автоматизации извлечения информации, предполагает наличие технологически развитого инструмента, способного выделять из текста нужные элементы, используя методы Text Mining. Работа данного инструмента состоит в том, чтобы структурировать подающуюся на вход текстовую информацию. Анализировать подобные наборы данных значительно проще и быстрее, чем результаты работы поисковика. На данном этапе также имеет смысл применение средств Text Mining – таких как реферирование, классификация, кластеризация – в зависимости от поставленной цели.

Следующий подход базируется на автоматизации аналитических процедур. В простейшем случае в руках конечного пользователя есть такие аналитические инструменты, как например, MS Excel и MS Access или в усовершенствованном варианте – BI и Data Mining. С целью оптимизации процесса обработки неструктурированной информации напрашивается очевидное решение: совместить инструменты Text Mining с инструментами анализа. Интегрировать элементы между собой можно при помощи базы данных. Для автоматизации процесса нужен некий механизм, который запросит информацию у ИПС, обнаружит искомые факты, структурирует их, сохранит в базе и сообщит о выполненном задании. Тогда аналитик должен будет только открыть отчеты и проанализировать результаты.

Технологии добычи информации из неструктурированных текстов (Text Mining) используются на практике уже сегодня. Со временем их применение будет только расширяться, поскольку объемы доступной и полезной информации растут с каждым днем, а потребность в их анализе по-прежнему не удовлетворена. В данной выпускной квалификационной работе будет реализован второй рассмотренный подход – подход, основанный на автоматизации извлечения информации, поскольку именно легковесные и привычные рядовым пользователям инструменты по обработке и структуризации информации наиболее востребованы на данный момент.

**1.2 Анализ существующих решений.**

Рынок инструментов Data Mining и Text Mining представлен множеством инструментов. Разработкой в данном секторе всемирного рынка программного обеспечения заняты как известные лидеры, так и новые развивающиеся компании. Инструменты могут быть представлены либо как самостоятельное приложение, либо как дополнения к основному продукту. Как правило, это масштабируемые системы, в которых реализованы различные математические и лингвистические алгоритмы анализа текстовых данных, имеющие развитые графические интерфейсы, богатые возможности визуализации и манипулирования данными, предоставляющие доступ к различным источникам данных. Примерами таких систем являются:

• Intelligent Miner for Text (IBM);

• TextAnalyst (Мегапьютер Интеллидженс);

• Text Miner (SAS);

• SemioMap (Semio Corporation);

• Knowledge Server (Autonomy);

• Galaktika-ZOOM (корпорация «Галактика»);

• InfoStream (Информационный центр «ЭЛВИСТИ»).

Рассмотрим некоторые из этих систем более подробно:

1. IBM Intelligent Miner for Text – это инструментарий для разработки программного обеспечения, предназначенного для извлечения знаний. Он содержит инструменты для программистов приложений, которые хотят создать приложения для извлечения ключевой информации из очень большого количества документов, электронных писем или веб-страниц, хранящихся в Интернете или в Интрасетях, без необходимости их читать. С помощью Intelligent Miner for Text можно осуществлять следующие задачи:

• Сортировка документов по темам, выделение преобладающих в коллекции тем и аннотация тем и документов;

• Поиск релевантных документов с использованием мощных и гибких инструментов построения запросов. 14 Intelligent Miner for Text содержит четыре основных компонента:

• IBM Text Analysis Tools: включают инструмент идентификации языка, комплексные инструменты кластеризации, инструмент аннотирования и средства извлечения ключевых понятий. Эти инструменты определяют язык документа, групповые концептуально связанные документы, классифицируют документы по контенту, выводят сводки документов и извлекают ключевые элементы текста.

• IBM Text Search Engine: многофункциональная поисковая система, которая настраивается как для сложного полнотекстового поиска (включая функции интеллектуального анализа текста), так и для веб-поиска. Поисковая система дополняется шаблонами Java (TM) и Java Beans, которые помогают создавать приложения для текстового поиска и административные функции, доступные из браузера с поддержкой Java.

• Пакет IBM Web Crawler: состоит из готового к запуску поискового робота и инструментария Web Crawler для создания настраиваемых поисковых роботов.

• IBM NetQuestion Solution – решение для текстового поиска на локальном Web-сайте или на нескольких Интернет или Интранет-серверах, основанное на Text Search Engine и Web Crawler. Стоимость продуктов разных уровней семейства Intelligent Miner составляет от 18 до 75 тысяч долларов. Согласно статистике, приведенной в источнике Intelligent Miner for Text чаще остальных представленных продуктов используется в корпоративной среде, где необходимо применение интеллектуального анализа текстов.

1. TextAnalyst был разработан российской компаний «Мегапьютер Интеллидженс» в качестве инструмента для анализа содержания текстов, смыслового поиска информации, формирования электронных архивов, и предоставляет пользователю следующие основные возможности:

• анализа содержания текста с автоматическим формированием семантической сети с гиперссылками – получения смыслового портрета текста в терминах основных понятий и их смысловых связей;

• анализа содержания текста с автоматическим формированием тематического древа с гиперссылками – выявления семантической структуры текста в виде иерархии тем и подтем;

• смыслового поиска с учетом скрытых смысловых связей слов запроса со словами текста;

• автоматического реферирования текста – формирования его смыслового портрета в терминах наиболее информативных фраз;

• кластеризации информации – анализа распределения материала текстов по тематическим классам;

• автоматической индексации текста с преобразованием в гипертекст;

• ранжирования всех видов информации о семантике текста по «степени значимости» с возможностью варьирования детальности ее исследования;

• автоматического/автоматизированного формирования полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации. Система TextAnalyst предоставляется методы Text Mining в качестве отдельного математического аппарата, который разработчики программного обеспечения могут встраивать в свои продукты, не опираясь на платформы информационно-поисковых систем или СУБД. Основная платформа для применения системы – MS Windows 9x/2000/NT. Существует также плагин TextAnalyst для браузера Microsoft Internet Explorer.

3. Американская компания SAS Institute выпустила систему SAS Text Miner для сравнения определенных грамматических и словесных рядов в письменной речи. Text Miner – весьма универсален, поскольку может работать с текстовыми документами различных форматов – в базах данных, файловых системах и даже в Web. Text Miner обеспечивает логическую обработку текста в среде мощного пакета SAS Enterprise Miner. Это позволяет пользователям обогащать процесс анализа 16 данных, интегрируя неструктурированную текстовую информацию с существующими структурированными данными. Основными преимуществами Text Miner являются:

* Улучшение модели данных: предиктивные модели использую ситуационные знания для описания будущих сценариев. Однако, важные обстоятельства и события, описанные в полях комментариев, заметках, отчетах, запросах, веб-комментариях и т. д., не попадают в структурные поля модели, несмотря на то, что данные достаточно легко анализируются. С помощью Text Miner появляется возможность добавлять выделенные из текстовых источников идеи в свои модели для большей прогностической способности.
* Автоматическое обучение: автоматизируются трудоемкие ручные операции, такие как извлечение темы или ключевых терминов, путем использования машинного обучение и методов обработки естественного языка. Высокопроизводительные процедуры позволяют получать результаты за считанные минуты, даже для больших коллекций.
* Добавление тематического исследования: Text Miner позволяет направлять результаты машинного обучения с помощью интерактивных графических интерфейсов, чтобы легко идентифицировать релевантность, модифицировать алгоритмы, принадлежность документов и групповые материалы в значимые совокупности. Результаты исследований можно расширять за пределы базовых начальных и конечных тематических списков, чтобы уточнить автоматически созданные правила и темы.
* Вмешательство в алгоритмы обработки: текст структурирован в числовые представления, которые суммируют коллекции документов и становятся исходными данными для методов прогнозирования и интеллектуального анализа данных. Используя ту же визуальную среду, что и SAS Enterprise Miner, можно легко изучить ключевые темы, выявить наиболее значимые фразы и увидеть, как меняются условия с течением времени, поэтому пользователь всегда будет знать, что изменить или включить в алгоритм для достижения лучших результатов. 17 Вместе с тем, компания SAS отмечает, что продукт Text Miner предназначен, в основном, для применения к бизнес-задачам.

4. SemioMap – это продукт компании Entrieva, созданный в 1996 г. ученым-семиотиком Клодом Фогелем (Claude Vogel). В мае 1998 г. продукт был выпущен как промышленный комплекс SemioMap 2.0 – первая система Text Mining, работающая в архитектуре клиент-сервер. Система SemioMap состоит из двух основных компонент – сервера SemioMap и клиента SemioMap. Работа системы протекает в три фазы:

• Индексирование – сервер SemioMap автоматически читает массивы неструктурированного текста, извлекает ключевые фразы (понятия) и создает из них индекс;

• Кластеризация понятий – сервер SemioMap выявляет связи между извлеченными фразами и строит из них, на основе совместной встречаемости, лексическую сеть («понятийную карту»);

• Графическое отображение и навигация – визуализация карт связей обеспечивает быструю навигацию по ключевым фразам и связям между ними, а также возможность быстрого обращения к конкретным документам. SemioMap поддерживает разбиение материала по «папкам», создание отдельной базы данных для каждой папки. Связи между понятиями, которые выявляет SemioMap, базируются на совместной встречаемости фраз в абзацах исходного текстового массива. Центральным блоком SemioMap является лексический экстрактор – программа, которая извлекает фразы из текстовой совокупности и выявляет совместную встречаемость этих фраз (их взаимные связи). Лексический экстрактор базируется на патентованной технологии SEMIOLEX. Она реализует идеи вычислительной семиотики, науки о знаках в языковой коммуникации, разработанной Клодом Фогелем.

5. Программный комплекс Галактика-ZOOM предназначен для аналитической обработки текстовых неструктурированных документов, находящихся в подключаемых базах данных. 18 Галактика-ZOOM объединяет три технологии: классическая поисковая система, система сбора текстовых данных (Text Mining), система аналитической обработки информации. Программный комплекс Галактика-ZOOM позволяет распределять по тематике информационный поток, исследовать, анализировать и выявлять тенденции. В список задач, решаемых инструментом Галактика-ZOOM входит:

• поиск в большом объеме неструктурированной информации.

• анализ найденной информации;

• анализ изменений проблемы и интереса к ней во времени;

• информационная «разведка» – добыча уникальной разрозненной прямой и косвенной информации;

• формирование и ведение тематических досье с возможностью вскрытия тенденций;

• вскрытие и исследование потенциальных угроз и тревожных тенденций;

• анализ «негатива» – выявление источников, целей, заказчиков, причин и поводов;

• выявление и исследование «информационных аномалий» в массивах данных; • реферирование текстов;

• вскрытие взаимосвязей персон, событий, процессов, тенденций и их неявной корреляции;

• автоматизированное решение типовых маркетинговых задач, например, задач «рейтингования» нескольких предприятий в том или ином регионе. Система содержит конверторы часто встречающихся форматов: простой текст, RTF, DOC, HTML. Galaktika-ZOOM и функционирует в среде ОС Windows.

6. Охват, обобщение больших динамических информационных массивов, непрерывно генерируемых в Internet, требует качественно новых подходов. Ввиду чего возникает необходимость создания методов мониторинга информационных 19 ресурсов, тесно связанных с методологией контент-анализа – контент-мониторинга. Для получения качественных и количественных срезов такой мониторинг должен производится постоянно на протяжении не определенного заранее времени. Для решения этой задачи в Информационном центре «ЭЛВИСТИ» разработана технология InfoStream. Программно-технологические средства InfoStream включают три основные составляющих:

• центр сбора и обработки информации;

• центр организации интерактивного доступа к базам данных;

• центр контент-мониторинга.

Ядром механизма обработки контента InfoStream является полнотекстовая информационно-поисковая система InfoReS. Технология позволяет создавать полнотекстовые базы данных и осуществлять поиск информации, формировать тематические информационные каналы, автоматически рубрицировать информацию, гистограммы распределения весовых значений отдельных понятий, а также динамики их встречаемости по времени. Технология InfoStream позволяет обрабатывать данные в форматах MS WORD (DOC, RTF), PDF, и всех текстовых форматов (простой текст, HTML, XML). Системы на основе InfoStream в настоящее время функционируют на платформах таких ОС: FreeBDS, Linux, Solaris. Как можно видеть, рынок программного обеспечения Text Mining представлен множеством инструментов, на нем идет постоянная конкурентная борьба за потребителя. Такая конкуренция порождает новые качественные решения. Все большее число поставщиков стремятся объединить в своих инструментах как можно большее число современных методов и технологий. В то же время специалисты в работе отмечают отставание существующего программного обеспечения от теоретических разработок в связи со сложностью программной реализации некоторых новых теоретических разработок методов и алгоритмов Text Mining.

**1.3 Предложения по усовершенствованию существующих программных решений**