УДК 004.912
DOI: 10.22213/2410-9304-2020-1-72-82

Сокращение объема текстового документа на основе анализа его корреляционных зависимостей

С. В. Моченов, кандидат технических наук, профессор, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия
Р. Р. Ахметгалеев, аспирант, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия
С. А. Лазарев, студент, Московский физико-технический институт (национальный исследовательский университет), Москва, Россия

В статье рассматриваются вопросы анализа текстовой информации с целью сокращения ее объема и представления содержания текста произвольных размеров в виде реферата. Текст рассматривается как генеральная совокупность предложений. В качестве основы для проведения анализа текста используются частотные (весовые) характеристики слов, в частности, существительных, используемых автором при построении предложений. Определена роль отдельных категорий слов. На основе весовых характеристик все слова разделяются на многократно и однократно используемые. Сформулированы рекомендации по применению слов-фильтров для извлечения из текста определенных предложений или группы предложений и представления их пользователю. Разработана методика анализа текстового документа. Анализируемый текст разбивается на группы предложений. Многократные слова используются в качестве базовых слов при определении корреляционных зависимостей между предложениями текста. На основе корреляционных зависимостей по каждой группе определяется одно приоритетное предложение, отражающее смысловую составляющую участка текста, задаваемого группой. За счет разбиения на группы достигается сокращение объема текста. Общее число приоритетных предложений соответствует числу групп. Эти предложения могут быть использованы для формирования реферата и предоставляют исследователю (пользователю) адекватную и сжатую информацию о содержании анализируемого документа. В статье приводятся примеры анализа, определяются направления дальнейших исследований.

Ключевые слова: анализ текстовой информации, многократные слова, однократные слова, корреляционные зависимости, приоритетное предложение, смысловое содержание.

Введение

С развитием цифровых технологий, в частности Интернета, резко возрастают объемы текстовой информации, которые становятся доступны пользователям. Исследователь (пользователь) сталкивается с проблемой поиска той информации, которая может оказаться полезной для решения его частных задач. Так, в работе [1] предлагается методология исследований для выявления проблем, связанных с информационной деятельностью МЧС России. В работе [2] проблема информационного поиска рассматривается в рамках подготовки лекционного материала преподавателями, а в работе [3] анализируются итоги автоматической оценки содержания сводок новостей.

Значительное количество работ связано с методами обработки и извлечения из текстов полезной информации, основанными на различных методах анализа.

В работе [4] автоматизированное реферирование определяется как перспективное направление компьютерной обработки текстов. Предлагается методика реферирования, основанная на содержательном анализе всего исходного текста, что позволяет говорить об адекватности семантики исходного текста и реферата.

В работе [5] рассматривается метод автоматизации процесса реферирования, основанный на построении содержательной модели текста в сочетании с моделью предметной области.

В работе [6] рассматривается модель ранжирования важности слова на основе набора функций оценки важности, что повышает эффективность использования ключевых слов при анализе мультидокументов новостей.

В работе [7] текст рассматривается как структура взаимосвязанных понятий, обеспечивающих продвижение к конечной цели, выражаемой авторской идеей. Предложена модель вектора цели. Отмечается возможность использования векторной модели и технологии обработки текстовой информации на ее основе для анализа и синтеза документов, выполнения направленного поиска и автоматического реферирования.

В работе [8] даются определения таких основных понятий, как аннотация, реферат и др. Сформулированы требования к содержательной стороне этих понятий для их применения в онлайновой поисковой системе.

В работе [9] рассматриваются особенности машинного перевода, целью которого является перевод текста с языка оригинала на целевой язык. Для повышения качества перевода рас-

сматриваются модели анализа текста, учитывающие зависимости как между отдельными словами, так и между отдельными фразами.

В работе [10] проводится подробный анализ организации синтаксических парсеров. Отмечается, что качество синтаксического анализатора во многом определяет качество решения задачи, выполняемой системой анализа текста. Указывается на трудности построения синтаксического парсера для русского языка и ошибки, возникающие при анализе сложных предложений. Дается описание авторской системы семантико-синтаксического анализа предложений русского и английского языка. Система позволяет выделить предикатные структуры предложений и построить деревья синтаксического подчинения предложений.

Представленный краткий обзор методов поиска и анализа текстовой информации позволяет сделать вывод об актуальности задач, связанных с данным направлением практической деятельности.

Автоматизация процессов поиска и анализа текстовой информации позволяет сократить время получения исследователем необходимых данных и ускорить решение стоящих перед ним задач. Каждый автор какого-либо текста, научной статьи в процессе написания опирается на собственную систему знаний, словарный запас и оперирует терминами, опираясь на грамматические основы языка. В свою очередь, пользователь (исследователь) также опирается на свою систему знаний, которая может существенно отличаться от системы знаний автора текста.

Структура поискового запроса зависит от целей организации поиска. При создании коллекции документов запрос строится на основе представлений исследователя в виде обобщенного образа нужной ему информации. В качестве такого обобщенного образа выступают ключевые слова, их комбинации или смысловые группы слов, сформулированные исследователем. Результаты поиска зависят как от алгоритмов работы поисковой системы, так и от качества поискового запроса.

При обработке коллекции документов главной задачей становится анализ отдельных документов и выявление в них полезной для исследователя информации.

При анализе отдельного текстового документа отсутствие предварительной информации о содержании текста, конкретных деталях и разница в используемых базах знаний автора и исследователя не дает возможности исследователю правильно построить смысловые фильтры для выделения из текста полезной информации и отбраковки второстепенной. Поисковый запрос или смысловой фильтр целесообразно строить на основе предварительных знаний о содержании текста. В свою очередь, содержание текста сформировано автором на основе его собственной системы знаний, которая в определенной мере косвенно представлена в тексте документа.

Актуальным становится вопрос о предварительной автоматической обработке текста и предоставлении пользователю полноценных сведений о содержании документа. На основании этих сведений пользователь может либо сделать вывод об отсутствии в документе нужной информации, либо использовать эти сведения для оптимального построения поискового запроса или смыслового фильтра и более глубокого анализа документа. Основой для такого анализа является информация, заложенная автором в тексте, ассоциативно связанная с системой знаний пользователя и вопросом, на который он ищет ответ.

Не менее актуальным является вопрос сокращения объема текстового документа при сохранении его смыслового содержания. Это особенно важно при построении: ассоциативных баз знаний по конкретным научным направлениям; систем обработки, кодирования и передачи данных; систем автоматизированного документооборота; для формирования реферата или аннотации документа и др.

Описание методов исследования

В работе [11] предложено использовать частотные характеристики слов, используемых в тексте или в его фрагменте, для определения темы документа.

В работе [12] на основе частоты использования в тексте слов (существительных) определяются веса слов и формируются весовые характеристики предложений. На основе весовых характеристик предложений осуществляется фильтрация текста и выбираются предложения, соответствующие выбранному пользователем значению веса. Возможно сокращение текста до объема, выражаемого в процентах, определенного пользователем.

В работе [13] описана система и алгоритм ее функционирования, основанные на использовании смысловых фильтров для выбора предложений текста, соответствующих критерию поиска, задаваемого смысловым фильтром.

В названных работах параметры фильтрации выбираются пользователем самостоятельно, без учета смысловой составляющей текстового документа, что ограничивает применимость подобных систем при обработке больших массивов текстовой информации. Целесообразно автоматизировать процедуру выбора параметров фильтрации.

В работе [14] формирование смысловых групп основано на построении базы знаний допустимых комбинаций частей речи для русскоязычных предложений. Это позволяет избежать сложных процедур семантического анализа при разбиении сложного предложения на смысловые группы.

В данной статье предлагаются варианты улучшения качества анализа и сокращения объема анализируемой информации при сохранении основной смысловой составляющей текстового документа. Новизна работы заключается в методике выбора приоритетных предложений, основанной на приоритете базовых слов и определении корреляционных зависимостей между предложениями текста.

Текст рассматривается как генеральная совокупность элементов, в качестве которых могут выступать слова предложений, сами предложения, группы предложений (например, абзацы), смысловые группы предложений.

Наиболее понятной является интерпретация какого-либо текста в виде генеральной совокупности отдельных слов. В работе [15] приводятся данные о распределении частей речи для русскоязычных текстов. Отмечается, что ведущая роль принадлежит существительным, прилагательным и глаголам.

Распределение конкретных слов, в частности существительных, по частоте их использования в тексте дает общее представление о содержании текста. Более детальный анализ может быть основан на выявлении роли отдельных категорий слов (существительных), составляющих основу предложений текста.

Все существительные (в дальнейшем – слова) в предложениях текста можно разбить на две основные категории: 1) слова многократного использования (многократные), 2) слова однократного использования (однократные).

Многократные слова также можно разбить на две группы.

1. Слова с высокой частотой использования в тексте или его части по сравнению с другими словами. Это группа, состоящая из двух, максимум трех слов, определяющих основную тему текста. Эти слова корреляционно связывают между собой предложения текста, позволяя пользователю рассматривать содержание текста или его части как единое целое.

2. Слова с меньшей частотой использования по сравнению со словами первой группы. Это группа разных слов, распределенных по отдельным предложениям или по группе соседних предложений, объединяющих особенности смысловой нагрузки этих предложений в единое целое, подтему основной темы. Это слова, которые входят в состав отдельных предложений и вносят соответствующую долю в смысловую составляющую рассматриваемого участка текста или всего текста.

Однократные слова используются в тексте или его части один раз и имеют вес, равный единице, если подсчет весов производился, соответственно, по всему тексту или его части. Эти слова входят в состав только разных предложений и совместно с другими словами придают определенное смысловое содержание конкретному предложению.

Перечисленные категории слов входят в состав смысловых групп, определяющих тему и рему предложений, т. е. связаны с ответами на вопросы «о чем говорится?» и «что говорится?». Слова первой категории, используемые в тексте многократно, распределены по предложениям текста или его части.

Между отдельными предложениями или отдельными частями текста существуют определенные смысловые зависимости (корреляционные связи), заложенные автором текста. С позиций статистики корреляция предполагает определение количественных отношений (степень связи) между элементами некоторой генеральной совокупности. Выявление этих связей позволит выделить приоритетные предложения, связанные между собой по смыслу и отражающие с достаточной полнотой общее содержание текста.

Слова и их комбинации (смысловые группы) в предложениях определяют смысловую составляющую предложений. Смысловая составляющая текста складывается из смысловых составляющих предложений и формируется и изменяется автором по ходу текста. Поэтому для выявления наиболее приоритетных предложений, относящихся к определенным аспектам смысловой составляющей текста, необходимо учитывать корреляционные зависимости между различными частями текста. Одновременно выделение приоритетных предложений позволит сократить объем анализируемого текстового документа.

В качестве меры связи между предложениями в данной работе используются весовые характеристики слов (существительных). Вес того или иного слова, его кратность определяются числом вхождений словоформ этого слова в предложения всего текста или его анализируемой части. Чем больше кратность слова, тем больше вероятность его присутствия в предложениях текста и выше его приоритет при определении наиболее значимых предложений, определяющих основную смысловую составляющую текста.

Предлагаемая в данной работе методика анализа текстового документа включает в себя следующие основные этапы.

1. Этап определения весов и приоритета многократно используемых слов (кратность слов).

2. Этап разбиения текста на группы предложений.

3. Этап определения корреляционных зависимостей предложений по многократным словам.

4. Этап сокращения объема текста и формирования массива наиболее значимых предложений.

На первом этапе осуществляется вычисление весов многократных слов текста или его части.

На втором этапе текст разбивается на группы предложений, исходя из требований пользователя на объем предоставляемой ему информации.

На третьем этапе по каждому многократному слову устанавливается его корреляционная связь со всеми предложениями текста путем определения наличия или отсутствия словоформ многократного слова в соответствующем предложении.

На четвертом этапе на основе приоритетов многократных слов по каждой группе предложений выделятся одно приоритетное предложение. Число приоритетных предложений определяется числом групп, на которые был разбит анализируемый текст.

Сформированный таким образом массив приоритетных предложений отражает основную семантическую составляющую текста при одновременном сокращении его объема. Этот массив может быть использован, например, в качестве основы для реферата.

В реальной практике для анализа может быть выбран участок текста произвольных размеров и разного содержания. Автоматизация процесса анализа текста на основе предлагаемой методики позволит пользователю существенно сократить время на поиск нужной информации для решения тех или иных задач.

Полученные в ходе экспериментального исследования данные и их интерпретация

Значение корреляционной связи между предложениями текста связано с использованием в тексте слов различной кратности. На рис. 1 представлен исходный текст из восемнадцати предложений, взятый в качестве примера, а на рис. 2, 3 показано распределение многократных и однократных слов этого текста. Эксперименты проводились с использованием информационной системы, описанной в работе [Втюрин М. В., 2017].

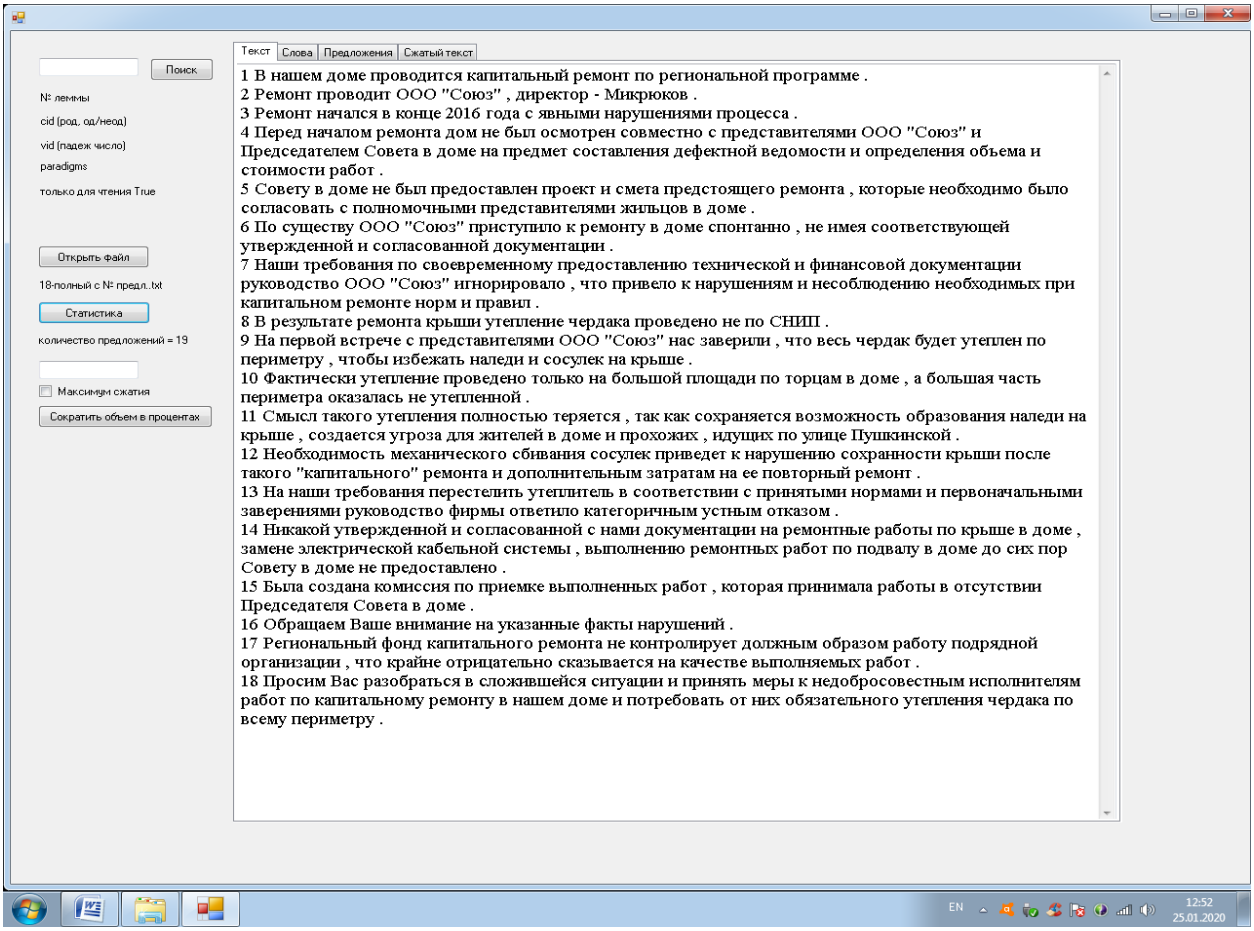

Рис. 1. Исходный текст

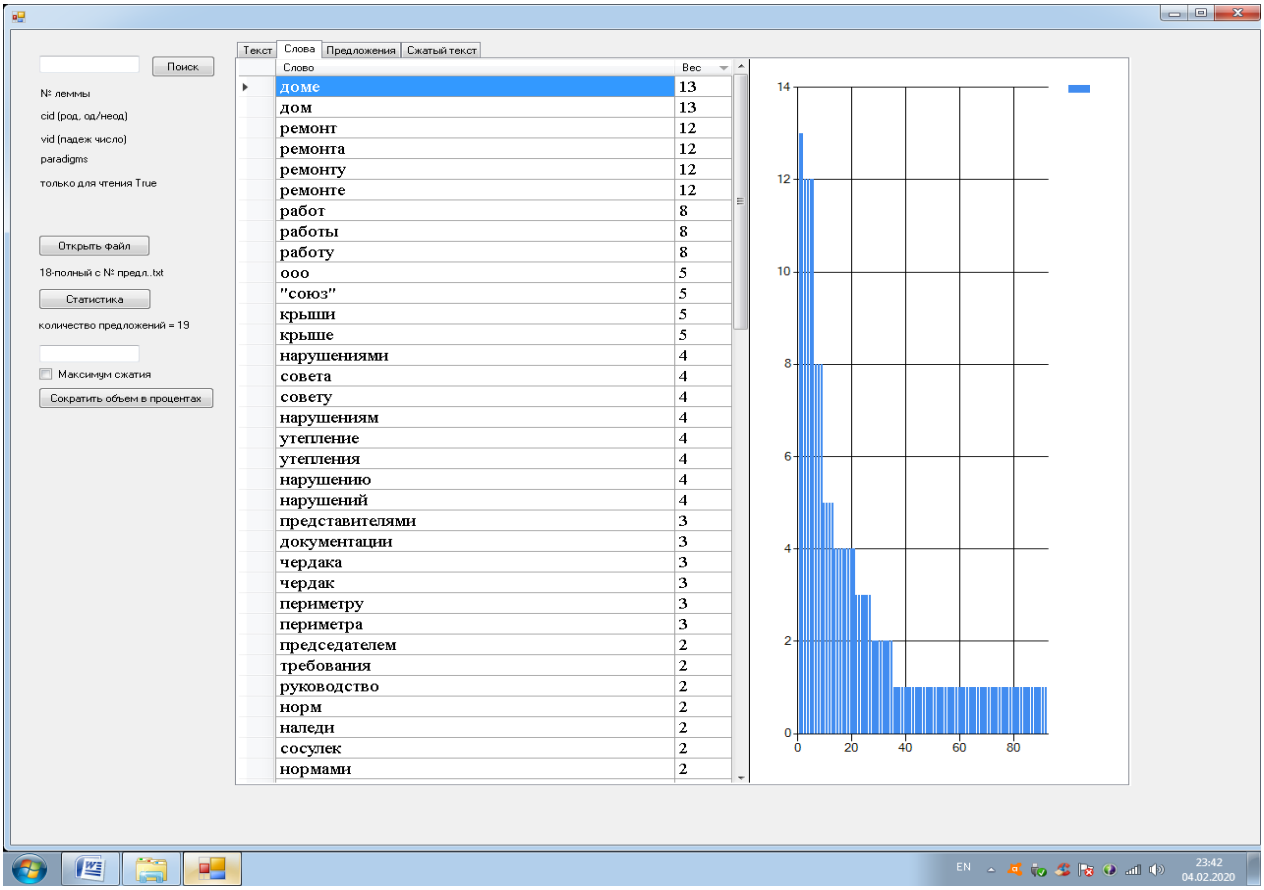

Рис. 2. Распределение многократных слов

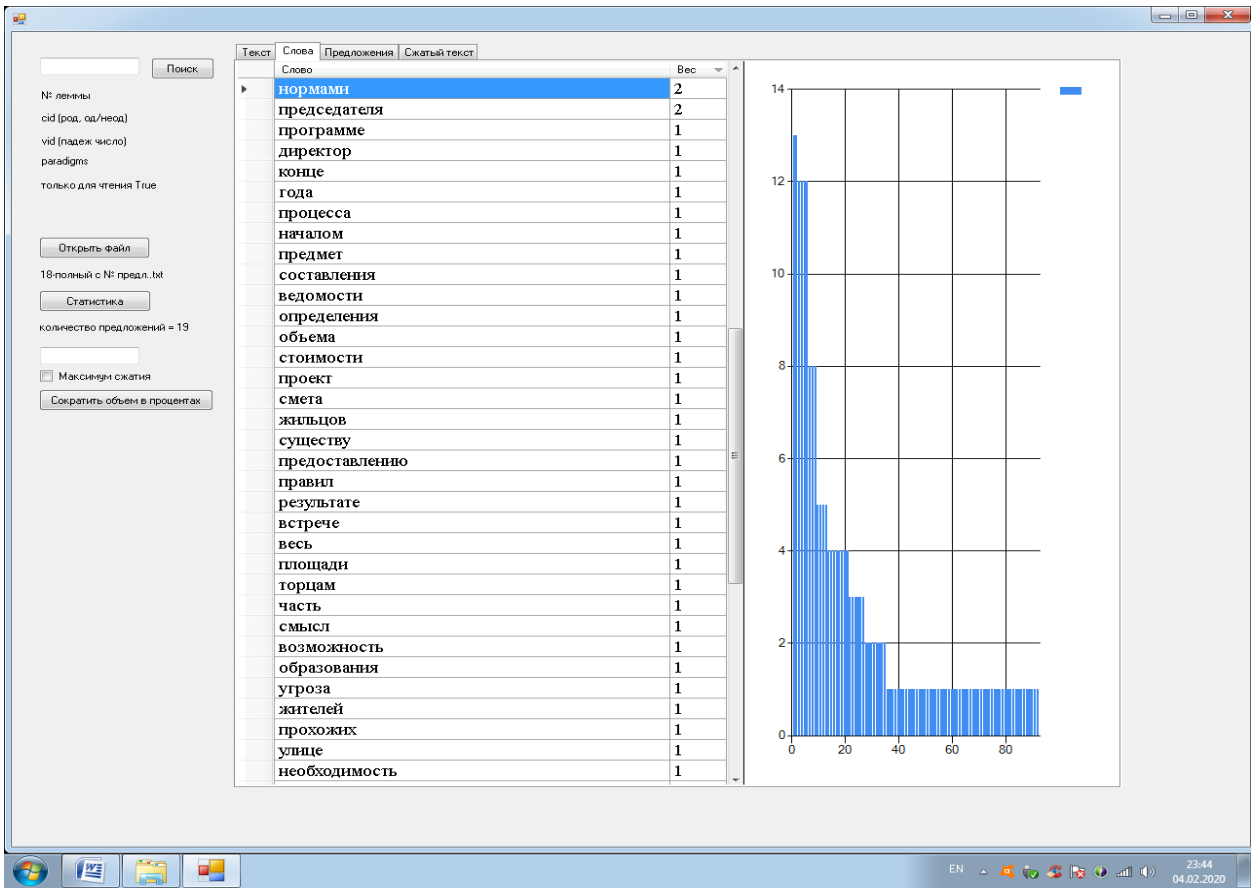

Рис. 3. Распределение однократных слов

Отмечается большое количество однократных слов – 41 % от общего числа слов (существительных). Это свидетельствует о существенной роли однократных слов при формировании автором смыслового содержания отдельных предложений и всего текста в целом.

На основании экспериментов, проведенных авторами данной статьи, можно сформулировать следующие рекомендации по применению слов-фильтров для извлечения из текста определенных предложений или группы предложений и представления их пользователю.

1. Однократно используемые слова-фильтры позволяют пользователю выделить отдельное предложение и оценить его смысловое содержание.

2. Двукратные слова-фильтры позволяют выделить либо одно предложение с повторяющимися словоформами слова-фильтра, либо два отдельных предложения в составе анализируемого текста.

3. Аналогично, трехкратные слова-фильтры позволяют выделить одно, два или максимум три предложения, имеющих корреляционные связи друг с другом.

4. Сочетание однократных, двукратных и более кратных слов-фильтров позволит выделить группу предложений, корреляционно связанных с выбранными словами-фильтрами.

5. Выбор слов-фильтров с большим весом не гарантирует существенного снижения объема текста из-за большой вероятности использования словоформ такого слова во многих предложениях текста.

6. Однократные слова-фильтры, взятые из разных частей текста, гарантируют выборку числа предложений, равную числу выбранных однократных слов-фильтров. Однако содержание выбранных предложений без учета корреляционных связей зависит от удачного выбора однократных слов-фильтров и не всегда позволяет уловить смысл текста, передаваемого только этими предложениями.

7. Выбор слов-фильтров должен быть связан с содержанием передаваемой в тексте информации.

8. Двукратные и трехкратные слова-фильтры корреляционно связывают небольшую группу предложений (максимум из двух или трех предложений), а их комбинация позволяет выбрать до пяти предложений из всего текста.

9. Аналогичные выводы можно сделать и по отношению к словам-фильтрам большей кратности. В частности, сочетание пятикратных слов-фильтров с однократными словами-фильтрами позволяет выбрать максимум до шести предложений при удачном выборе однократного слова-фильтра, так как словоформы выбранного однократного слова могут содержаться в предложениях, выбранных по пятикратному слову-фильтру.

В общем случае для фильтрации могут быть использованы другие части речи и члены предложения, смысловые группы. Выбор того или иного варианта зависит от целей анализа, определяемых пользователем, исследователем. В данной статье эти вопросы не рассматриваются.

Перечисленные варианты фильтрации – это лишь часть возможных вариантов, основанных на использовании только существительных. Эти варианты фактически позволяют выбрать из текста определенное число предложений, но с возможным нарушением последовательности изложения, определенной автором. Более равномерный выбор предложений можно обеспечить за счет разбиения текста на группы из нескольких предложений, с последующим выбором в каждой группе одного приоритетного предложения для представления пользователю. Это позволит сократить объем текста и уменьшить риск потери смысловой связности между выбранными предложениями.

Для сокращения объема текста и сохранения его смыслового содержания вводятся следующие критерии.

1. Приоритетные предложения должны выбираться равномерно по тексту.

2. Количество предложений в выборке должно соответствовать запросам пользователя или общепринятым нормам (K), например, K=30 % от числа предложений всего текста для реферата.

3. Выбор приоритетного предложения в группе должен проводиться с учетом приоритета базовых слов и корреляционных связей между предложениями группы.

Следует отметить, что при числе предложений в группе более трех выбор приоритетного предложения существенно усложняется.

Для удовлетворения поставленным критериям весь текст из N предложений разбивается на группы по три предложения с целью определения в каждой группе одного приоритетного предложения. Количество групп определяется с учетом критерия 2. Например, при принятой норме K = 30 % по критерию 2 и общем числе предложений N = 101 число групп G будет равно $G=(N \cdot K)/100=30{,}3$. Если полученное значение G дробное, то оно округляется до ближайшего большего целого. При этом в последней группе может оказаться одно или два предложения.

В рассматриваемом примере для текста из восемнадцати предложений число групп $G = (18 \cdot 30)/100 = 5{,}40$. С учетом округления $G = 6$. В каждой группе получается по три предложения, из которых необходимо выбрать одно наиболее приоритетное. При увеличении нормы K до 50 % количество групп увеличивается до 9.

В таблице показано разбиение анализируемого текста на группы по три предложения в группе и распределение базовых слов по предложениям.

Разбиение текста на группы и распределение базовых слов по предложениям

Базовое слово	Группа 1 Предложения 1 2 3	Группа 2 Предложения 4 5 6	Группа 3 Предложения 7 8 9	Группа 4 Предложения 10 11 12	Группа 5 Предложения 13 14 15	Группа 6 Предложения 16 17 18	Кратность слова
дом	1 0 0	2 2 1		1 1 0	0 3 1	0 0 1	13
ремонт	1 1 1	1 1 1	1 1 0	0 0 2		0 1 1	12
работы		1 0 0			0 2 2	0 2 1	8
ооо	0 1 0	1 0 1	1 0 1				5
«союз»	0 1 0	1 0 1	1 0 1				5
крыши			0 1 1	0 1 1	0 1 0		5
нарушениями	0 0 1		1 0 0	0 0 1		1 0 0	4
совета		1 1 0			0 1 1		4
утепление			0 1 0	1 1 0		0 0 1	4
представителями		1 1 0	0 0 1				3
документации		0 0 1	1 0 0		0 1 0		3
чердака			0 1 1			0 0 1	3
периметру			0 0 1	1 0 0		0 0 1	3
председателем		1 0 0			0 0 1		2
требования			1 0 0		1 0 0		2
руководство			1 0 0		1 0 0		2
норм			1 0 0		1 0 0		2
сосулек			0 0 1	0 0 1			2
программе	1 0 0						1
директор	0 1 0						1
конце	0 0 1						1
года	0 0 1						1
процесса	0 0 1						1
началом		1 0 0					1

Как уже отмечалось выше, смысл предложения передается набором и последовательностью семантически связанных слов этого предложения. Наличие словоформ определенного базового слова в различных предложениях корреляционно связывает эти предложения между собой, при этом можно говорить о пространственной связности и степени связи между предложениями.

В приведенной таблице по выбранному базовому слову с определенной кратностью каждая строка определяет корреляционные связи между предложениями текста.

Значение корреляционной связи по отдельному предложению (степень связи) в каждой группе определяется числом использованных в данном предложении словоформ, соответствующих базовому слову. Нулевое значение говорит об отсутствии корреляционной связи данного предложения с другими предложениями по выбранному базовому слову. Пустые ячейки также соответствуют нулевому значению корреляционной связи и не заполнены нулями для большей наглядности таблицы.

Анализ таблицы по строкам, соответствующим большинству базовых слов, позволяет сделать вывод о том, что число корреляционно связанных между собой предложений зависит от кратности базового слова. Это подтверждает выше сделанные выводы по применению словфильтров. Чем больше кратность базового слова, тем больше вероятность образования цельного множества соседних предложений, корреляционно связанных между собой. Такое множество может включать в себя два и более предложения, формирующих смысловое содержание соответствующей части текста.

Предложения, имеющие в своем составе словоформу, связанную с базовым словом, распределены по группам неравномерно. Так, по базовому слову «ремонт» образовано компактное множество корреляционно связанных между собой предложений 1–8, входящих в группы 1,

2, 3. В то же время предложение 12 не входит ни в какое другое компактное множество и не имеет корреляционно связанных с ним соседних предложений, входящих в группу 4. Предложение 12 определяется как приоритетное, поскольку оно корреляционно связано с другими предложениями по базовому слову «ремонт», отличается по смысловому содержанию от соседних предложений текста и вносит в смысловое содержание текста дополнительную информацию через однократные слова.

Таким образом, выбор приоритетного предложения в той или иной группе ведется в порядке приоритета базовых слов. Из группы выбирается предложение, корреляционно связанное с базовым словом и не имеющее корреляционных связей с соседними предложениями группы.

При невозможности выбора приоритетного предложения по группе, в которой располагается несколько предложений (два или три), корреляционно связанных с базовым словом, осуществляется переход к следующему базовому слову с меньшей кратностью и проводится аналогичный анализ по выбранной группе.

Таким образом, в результате анализа корреляционных зависимостей по группам в выборку приоритетных предложений попадают:

1) по базовому слову «дом» – предложения 1 и 18;

2) по базовому слову «ремонт» – предложение 12;

3) по базовому слову «работы» – предложение 4;

4) по базовому слову «крыши» – предложение 14;

5) по базовому слову «нарушениями» – предложение 7.

На рис. 4 показан сокращенный текст из шести приоритетных предложений. Сформированная выборка приоритетных предложений включает в себя шесть предложений: 1, 4, 7, 12, 14, 18. При этом обеспечивается сокращение объема текста на 67 % (по числу предложений) и сохраняется основное смысловое содержание исходного текста.

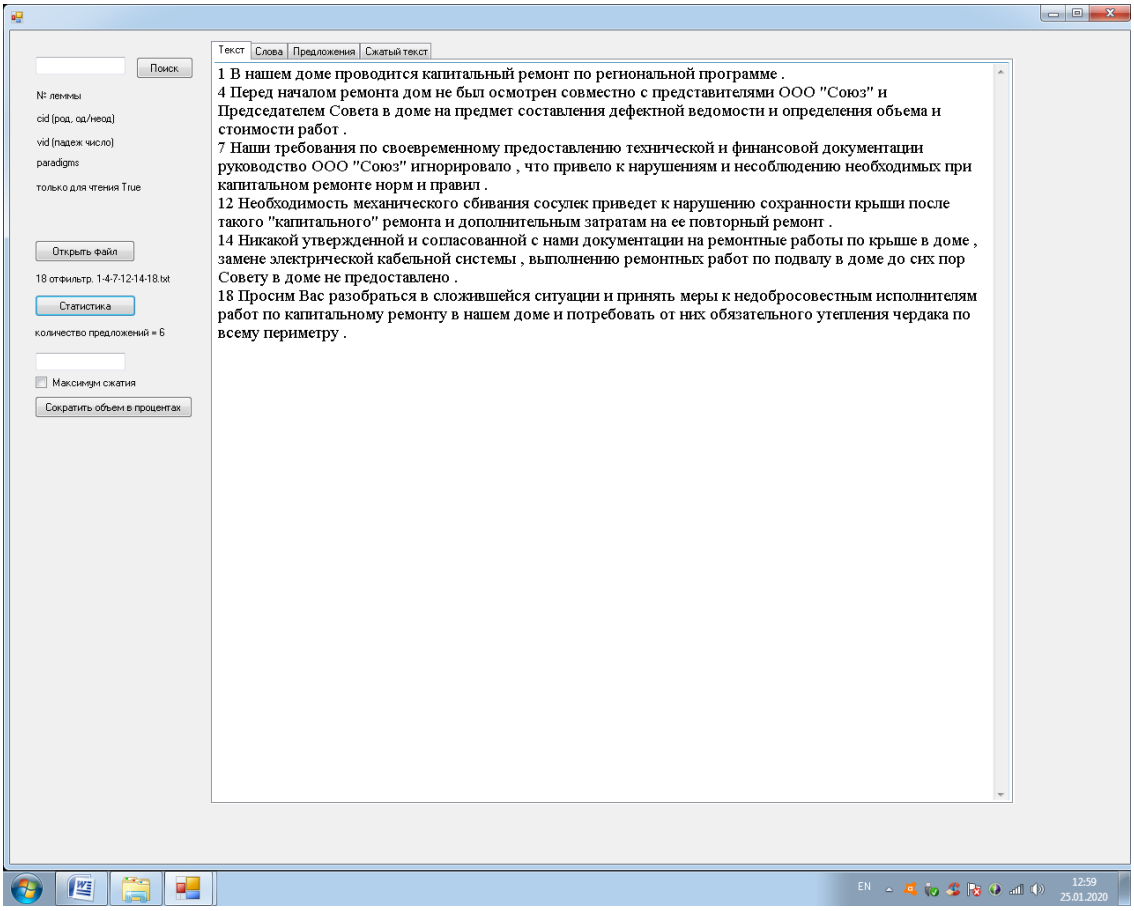

Рис. 4. Сокращенный текст из шести приоритетных предложений

Эти предложения могут быть использованы для построения реферата и дополнительного целенаправленного исследования текстового документа пользователем при решении его конкретных задач.

Заключение

Представленная методика и результаты ее применения показывают принципиальную возможность создания автоматизированной системы для получения сжатой и адекватной информации о содержании текстового документа. Это существенно сократит время, затрачиваемое исследователем, на поиск нужной информации.

В статье введено понятие многократных и однократных слов, определены возможности их использования в качестве слов-фильтров.

Разработана методика анализа текста, основанная на приоритете базовых (многократных) слов и определении корреляционных зависимостей между предложениями текста. Введены критерии анализа текста, которые обеспечивают равномерный выбор из текста приоритетных предложений и уменьшают риск потери смысловой связности между ними.

Выборка приоритетных предложений представляется исследователю и может быть использована для формирования реферата. Детали методики подробно рассмотрены на приведенном примере анализа текстового документа и иллюстрируются соответствующей таблицей. В дальнейшем предполагается расширение функционала используемой информационной системы и реализация ее прототипа на основе нейронных сетей.

Библиографические ссылки

1. *Артюхин В. В., Чяснавичюс Ю. К.* Планирование аналитического исследования при помощи методов анализа качественных данных // Прикладная информатика. 2014. № 2. С 23–48.

2. *Волкова Е. С., Моченов С. В., Шаронов М. А.* Проблема информационного поиска в педагогической практике // Вестник Ижевского государственного технического университета. 2014. № 4. С. 180-182.

3. *Rankel P., Conroy J., Dang H., Nenkova A.* A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013. pp. 131-136.

4. *Герте Н. А., Нестерова Н. М.* Реферирование как способ извлечения и представления основного содержания текста // Вестник Пермского университета. Российская и зарубежная филология. 2013. №4/24. С. 127–132.

5. *Курушин Д. С., Нестерова Н. М., Овчинникова И. Г.* О возможном подходе к созданию системы автоматического реферирования // Вопросы психолингвистики. 2014. № 2 (20). С. 123–128.

6. *K. Hong and A. Nenkova.* "Improving the Estimation of Word Importance for News Multi-Document Summarization," in EACL, 2014, pp. 712-721. URL: https://repository.upenn.edu/cgi/viewcontent.cgi?article=2036&context=cis_reports (дата обращения: 29.01.2020).

7. *Моченов С. В., Бледнов А. М., Луговских Ю. А.* Векторная модель представления текстовой информации // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам : материалы Междунар. науч. конф. (Ижевск, 13–17 июля 2006 г.) / отв. ред. В. А. Баранов. Ижевск : Изд-во ИжГТУ, 2006. С. 131–139.

8. [Abstracts - The Writing Center] [Электронный ресурс]. URL: http://writingcenter.unc.edu/handouts/abstracts/ (дата обращения 03.02.2020).

9. *Och F.J., Tillmann C., Ney H.* Improved Alignment Models for Statistical Machine Translation. URL: https://www.researchgate.net/publication/2282249_Improved Alignment Models for Statistical Machine Translation (дата обращения 04.02.2020).

10. *Харламов А. А., Ермоленко Т. В., Дорохина Г. В.* Сравнительный анализ организации систем синтаксических парсеров // Инженерный вестник Дона : электронный научный журнал. 2013. № 4. URL: http://ivdon.ru/ru/magazine/archive/n4y2013/2015 (дата обращения: 04.02.2020).

11. *Luhn H.P.* The automatic creation of literature abstracts // IBM Journal of Research and Development. 1958. Vol. 2, № 2. P. 159–165. URL: https://www.google.com/search?q=H.+P.+Luhn.+1958.+The+automatic+creation+of+literature+abstracts.+IBM+Journal+of+Research+and+Development (дата обращения: 29.01 2020).

12. *Втюрин М. В., Ястребов А. И., Моченов С. В.* Разработка информационной системы для уменьшения объема текстовой информации в процессе информационного поиска // Интеллектуальные системы в производстве. 2017. Т. 15. № 3. С. 94–99.

13. *Втюрин М. В., Моченов С. В.* Применение статистических характеристик для сокращения объема текстовой информации при сохранении ее информативности // Вестник ИжГТУ имени М.Т. Калашникова. 2018. Т. 21. № 2. С. 173–179.

14. *Моченов С. В., Ахметгалеев Р. Р.* Об одном подходе к построению информационной системы обработки текстовой информации на основе смысловых групп // Интеллектуальные системы в производстве. 2019. Т. 17. № 2. С. 58–64.

15. *Моченов С. В., Бледнов А. М., Луговских Ю. А.* Использование статистических методов для семантического анализа текста // Технологии информатизации профессиональной деятельности (в науке, образовании и промышленности) : сб. тр. науч.-техн. конф.и с междунар. участием в рамках форума «Высокие технологии – 2004». Ижевск : Регулярная и хаотическая динамика, 2005. С. 360–365.

References

1. Artjuhin V.V., Chjasnavichjus Ju.K. [Planning an analytical study using qualitative data analysis methods]. *Prikladnaja informatika.* 2014. No. 2. Pp. 23-48 (in Russ.).

2. Volkova Ye.S., Mochenov S.V., Sharonov M.A. [The problem of information retrieval in pedagogical practice]. *Vestnik Izhevskogo gosudarstvennogo tekhnicheskogo universiteta*, 2014. No. 4. Pp. 180-182 (in Russ.).

3. Rankel P., Conroy J., Dang H., Nenkova A. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013. Pp. 131-136.

4. Gerte N.A., Nesterova N.M. [Referencing as a way of extracting and presenting the main content of a text] // Bulletin of Perm University. Russian and foreign philology. 2013. No. 4/24. Pp. 127-132 (in Russ.).

5. Kurushin D.S., Nesterova N.M., *Ovchinnikova I.G.* [On a possible approach to the creation of an automatic abstracting system]. Issues of Psycholinguistics. 2014. No. 2. Pp. 123-128 (in Russ.).

6. Hong *K*. and Nenkova A. "Improving the Estimation of Word Importance for News Multi-Document Summarization," in EACL, 2014, pp. 712-721. URL: https://repository.upenn.edu/cgi/viewcontent.cgi?article=2036&context=cis_reports (accessed 29.01.2020).

7. Mochenov S.V., Blednov A.M., Lugovskikh Yu.A. *Vektornaja model' predstavlenija tekstovoj informacii* [Vector model of text information representation]. *Sovremennye informatsionnye tekhnologii i pis'mennoe nasledie: ot drevnikh rukopisei k elektronnym tekstam: international materials. scientific conf. (Izhevsk, July 13-17, 2006) / otv. ed. V.A. Baranov* [Proc. of Modern information technology and written heritage: from ancient manuscripts to electronic texts]. *Izhevsk, Izhevsk State Technical University publishing house*, 2006. Pp. 136-145 (in Russ.).

8. [Abstracts - The Writing Center] [Electronic resource]. URL: http://writingcenter.unc.edu/handouts/abstracts/ (accessed 03.02.2020).

9. Och F.J., Tillmann C., Ney H. Improved Alignment Models for Statistical Machine Translation. URL:https://www.researchgate.net/publication/2282249 Improved_Alignment_Models_for_Statistical_Machine_Translation (accessed 04.02.2020).

10. Kharlamov A.A., Yermolenko T.V., Dorokhina G.V. [Comparative analysis of the organization of syntactic parser systems]. Electronic Scientific Journal "Engineering Journal of the Don" 2013. No. 4. URL: http://ivdon.ru/ru/magazine/archive/n4y2013/2015 (accessed 04.02.2020).

11. Luhn H.P. The automatic creation of literature abstracts // IBM Journal of Research and Development. – 1958. – Vol. 2, № 2. – P. 159–165. URL:https://www.google.com/search?q=H.+P.+Luhn.+1958.+The+automatic+creation+of+literature+abstracts.+IBM+Journal+of+Research+and+Development (accessed 01.29.2020).

12. Vtyurin M.V., Yastrebov A.I., Mochenov S.V. [Development of an information system to reduce the amount of textual information in the process of information retrieval]. *Intellektualnye sistemy v proizvodstve*. 2017. Vol. 15. No. 3. Pp. 94-99. (in Russ.).

13. Vtyurin M.V., Mochenov S.V. *Primeneniye statisticheskikh kharakteristik dlya sokrashcheniya ob"yema tekstovoy informatsii pri sokhranenii yeye informativnosti* [The use of statistical characteristics to reduce the amount of textual information while maintaining its information content]. *Vestnik IzhGTU imeni M.T. Kalashnikova*. 2018. Vol. 21. No. 2. Pp. 173-179 (in Russ.).

14. Mochenov S.V., Akhmetgaleyev R.R. [On one approach to building an information system for processing text information based on semantic groups]. *Intellektualnye sistemy v proizvodstve.* 2019. Vol. 17. No. 2. Pp. 58-64 (in Russ.).

15. Mochenov S.V., Blednov A.M., Lugovskikh YU.A. *Ispol'zovaniye statisticheskikh metodov dlya semanticheskogo analiza teksta* [The use of statistical methods for semantic text analysis]. *Technologies of informatization of professional activity (in science, education and industry): Sat. tr scientific technology conferences with int. participation in the forum "High Technologies - 2004". Izhevsk: SRC "Regular and chaotic dynamics",* [Proc. of the Tehnologii informatizacii professionalnoj deyatelnosti (v nauke, obrazovanii i promyshlennosti) : sb. tr. nauch.-tehn. konf.i s mezhdunar. uchastiem v ramkah foruma «Vysokie tehnologii – 2004»]. 2005. Pp. 354-359 (in Russ.).

* * *

Reducing the Text Document Volume Based on Analysis of Its Correlation Dependencies

S. V. Mochenov, PhD in Engineering, Professor, Kalashnikov ISTU, Izhevsk, Russia
R. R. Ahmetgaleev, Post-graduate, Kalashnikov ISTU, Izhevsk, Russia
S. A. Lazarev, Student, Moscow Institute of Physics and Technology (National Research University), Moscow, Russia

The paper deals with the analysis of textual information with the aim of reducing its volume and presenting the content of text of arbitrary sizes in the form of an abstract. The text is considered as a totality of sentences. As a basis for text analysis, the frequency (weight) characteristics of words are used, in particular, nouns used by the author in constructing sentences. The role of certain categories of words is determined. Based on weight characteristics, all words are divided into repeatedly and once used. Recommendations are formulated on the use of filter words to extract certain sentences from a text or a group of sentences and present them to the user. A technique for analyzing a text document has been developed. The analyzed text is divided into groups of sentences. Multiple words are used as base words in determining correlation dependencies between sentences in a text. Based on the correlation dependencies for each group, one priority proposal is determined, which reflects the semantic component of the text section specified by the group. By splitting into groups, a reduction in text volume is achieved. The total number of priority proposals corresponds to the number of groups. These proposals can be used to form an abstract and provide the researcher (user) with adequate and concise information about the content of the analyzed document. The paper provides examples of analysis and identifies the areas for further research.

Keywords: analysis of textual information, multiple words, single words, correlation dependencies, priority sentence, semantic content.

Получено: 10.12.2020