УДК 004.912
DOI 10.22213/2413-1172-2018-2-173-179

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ДЛЯ СОКРАЩЕНИЯ ОБЪЕМА ТЕКСТОВОЙ ИНФОРМАЦИИ ПРИ СОХРАНЕНИИ ЕЕ ИНФОРМАТИВНОСТИ

М. В. Втюрин, аспирант, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия
С. В. Моченов, кандидат технических наук, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

Рассматривается возможность применения исследователями специализированных алгоритмов для информационной системы, обеспечивающей сокращение объема анализируемой текстовой информации в процессе информационного поиска. Актуальность работы обосновывается сложностью информационного поиска, связанного с решением пользователем конкретной задачи и необходимостью переработки больших объемов текстовых данных. Целью является сокращение объема анализируемой текстовой информации русскоязычных текстов при сохранении их смысловой составляющей.

Приведено описание ранее разработанной информационной системы для сокращения объема текстовой информации в процессе информационного поиска. Представлено описание двух различных подходов к анализу текста, что позволяет осуществить сравнительный анализ получаемых результатов. Выполнена реализация данных подходов на базе ранее разработанной информационной системы, в структурную схему и алгоритм функционирования которой внесены соответствующие изменения.

Приведены результаты проведенного экспериментального исследования. Из результатов применения описываемых подходов следует, что основная доля предложений, соответствующих запросу пользователя по выбранному тексту, представлена в заключительной части текста, что позволяет исследователю обратить внимание именно на эту часть анализируемого документа. Получены результаты, которые могут быть использованы для составления рефератов и аннотаций анализируемых документов. В дальнейшем предполагается формировать авторские смысловые группы слов, которые могут быть использованы исследователем для синтеза новых знаний.

Ключевые слова: анализ текстов, информационная система, текстовая информация, сокращение объема текста, информационный поиск.

Введение

Одним из приоритетных направлений исследования в области анализа текстов можно назвать анализ русскоязычных текстов. В частности, одной из проблем является возможность сокращения объема текстовой информации, анализируемой пользователем в процессе информационного поиска. Ранее этот вопрос рассматривался в работах [1–6].

Исследование является актуальным, поскольку с ростом информационных технологий растут объемы текстовой информации и возникает потребность в ее ускоренной обработке пользователями. Сокращение объема текстовой информации при сохранении ее информативности способствует уменьшению времени, затрачиваемого пользователем на проведение информационного поиска. Это, в свою очередь, позитивно влияет на эффективность работы пользователя.

В качестве предмета исследования в статье рассматриваются вопросы разработки алгоритмов сокращения объема текстовой информации при сохранении ее информативности.

Основными критериями для сокращения объема текстовой информации в данной работе являются статистические критерии оценки фрагментов текста (предложений, абзацев), основанные на повторяемости всех слов текста или только существительных.

Целью проводимого исследования является повышение эффективности работы пользователя при анализе текстовой информации в процессе информационного поиска. Для достижения этой цели были выделены следующие задачи: выполнить анализ существующих методов сокращения текстовой информации, разработать алгоритмы сокращения объема текстовой информации, разработать информационную систему, на базе которой можно будет апробировать разработанные алгоритмы.

Описание методов исследования

Одним из подходов к рассмотрению сокращения объема текстовой информации может быть способ, состоящий в оценивании фрагментов текста (предложений, абзацев) на основе статистики повторяемости слов текста. При

этом ожидается выделение наиболее информативных фрагментов текста и сохранение смысловой нагрузки исходной информации.

Лун Г. П. в своей работе [7] выдвинул идею о том, что некоторые слова документа описывают его содержимое, и предложения, передающие наиболее важную информацию в этом документе, содержат много таких описывающих слов, расположенных близко друг к другу. Он также предложил использовать частоту появления слов, чтобы определить, какие слова описывают тему документа. По его мнению, слова, которые часто встречаются в документе, скорей всего и будут являться темой этого документа.

Описание используемой теоретической модели

Ранее была разработана информационная система [8], выполняющая сокращение текстовой информации путем сравнения поискового запроса с текстом выбранного документа.

Структурная схема системы и алгоритм ее функционирования представлены на рис. 1, 2.

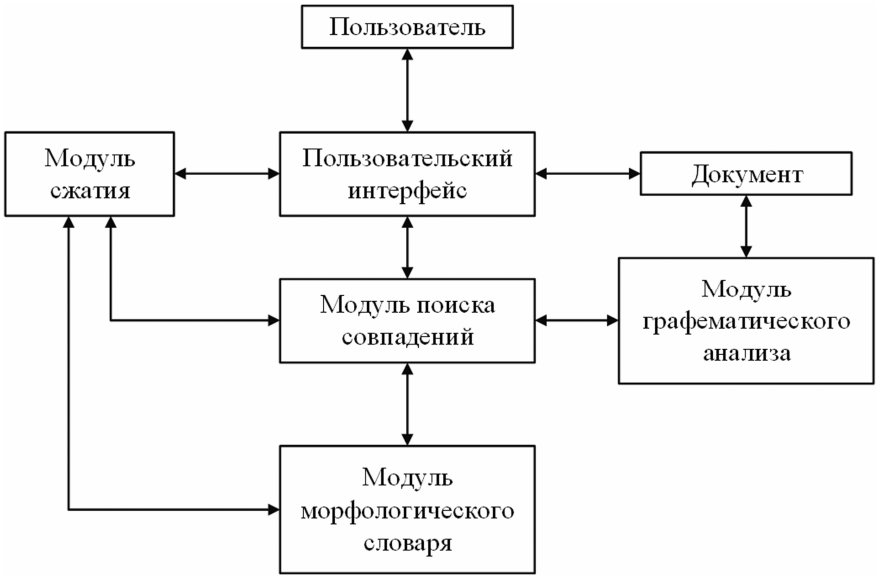

Рис. 1. Структурная схема информационной системы сокращения текстовой информации

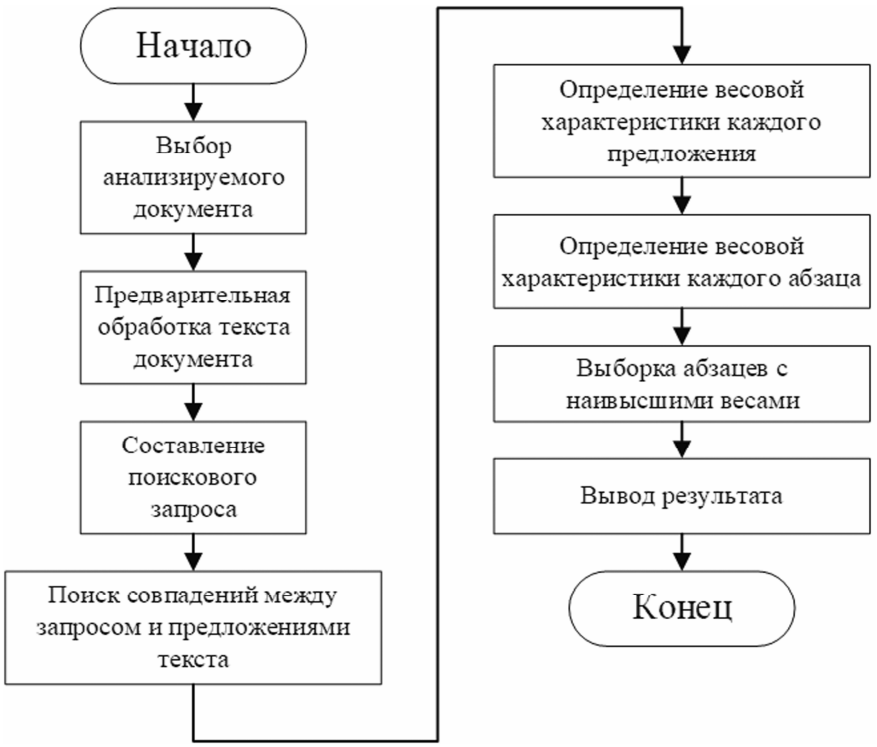

Рис. 2. Алгоритм функционирования информационной системы

Пользователь (исследователь) через пользовательский интерфейс информационной системы выбирает документ для анализа и вводит поисковый запрос. Поисковый запрос представляет собой одно или несколько словосочетаний, отражающих интересующую исследователя информацию.

Модуль графематического анализа подготавливает текст для дальнейшей работы, разбивая его на составляющие части: абзацы, предложения и слова.

Модуль морфологического словаря с помощью библиотеки mcr.dll [9, 10] выдает уникальный номер для каждого слова текста. На основе уникального номера модуль поиска совпадений выполняет сравнение поискового запроса со словами предложений исходного текста. Однокоренные слова получают один и тот же уникальный номер. В зависимости от количества совпавших по уникальному номеру слов модуль поиска присваивает количественную оценку каждому предложению. Данная оценка отображает степень соответствия пользовательского запроса анализируемому предложению.

Каждому абзацу присваивается наивысшая оценка, выбираемая из оценок предложений, входящих в этот абзац. Модуль сжатия выполняет выборку абзацев текста с наивысшими оценками. Выбранные абзацы представляются пользователю в качестве результата работы системы.

В новой версии разработанной информационной системы реализованы два различных подхода к анализу текста, что позволяет осуществить сравнительный анализ получаемых результатов. Первый подход основан на использовании частоты повторяемости всех слов текста, второй – на использовании только существительных.

Измененная структурная схема представлена на рис. 3.

Модули системы работают аналогично предыдущему варианту со следующими изменениями.

Модуль сжатия текста по-прежнему выполняет выборку элементов текста по их оценкам. Дополнительно используются результаты работы модуля статистики. Оценки соответствия предложения поисковому запросу обладают большим приоритетом, поэтому предложения, полностью соответствующие поисковому запросу (имеющие оценку 100 %), включаются в результат.

Для решения о том, стоит ли включать в результат предложение, частично соответствующее поисковому запросу, используется статистика повторяемости слов. Таким образом, формируются два варианта результатов. В первом варианте предложения, частично соответствующие запросу пользователя, оцениваются по статистике повторяемости всех слов текста, а во втором – по статистике повторяемости существительных.

Алгоритм функционирования измененной информационной системы представлен на рис. 4.

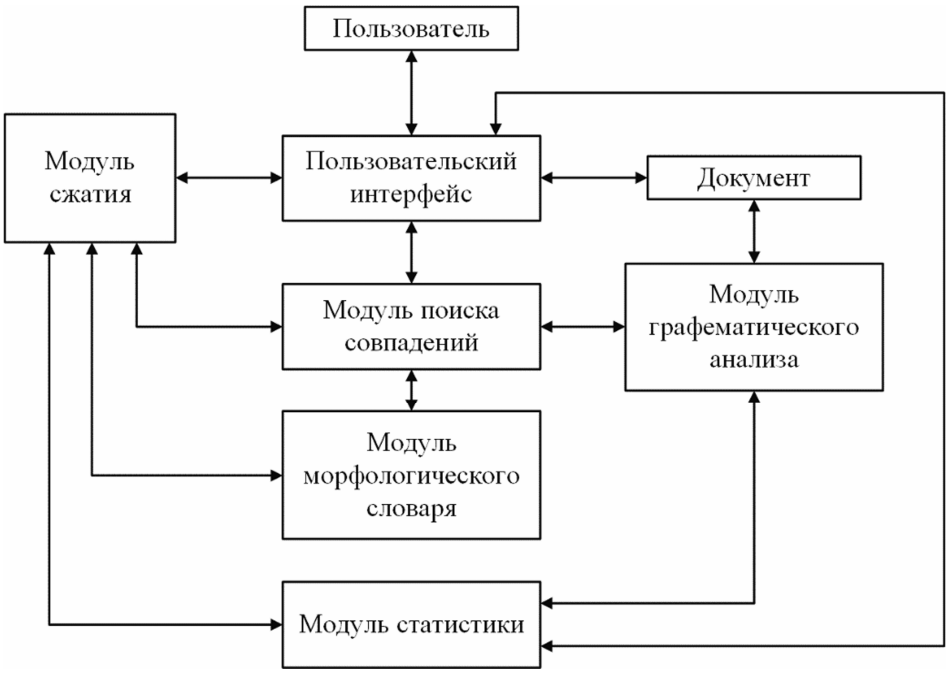

Рис. 3. Структурная схема информационной системы с использованием модуля статистики

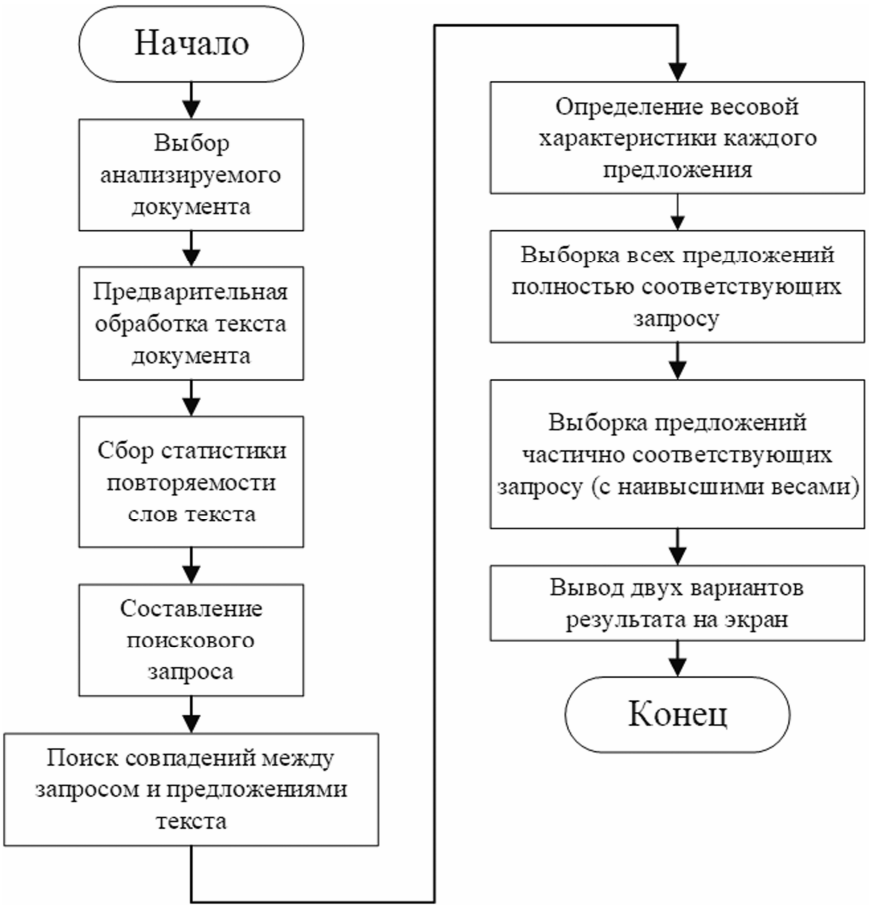

Рис. 4. Алгоритм работы системы с использованием статистики повторяемости слов

В описанной информационной системе используется словарь А. А. Зализняка. Программное обеспечение разработано в среде разработки MicrosoftVisualStudio 2017 Community, язык разработки – C#.

Для проведения экспериментального исследования был выбран текст научной статьи «Разработка информационной системы для уменьшения объема текстовой информации в процессе информационного поиска» (Втюрин М. В., Ястребов А. И., Моченов С. В.) объемом 745 слов. Был составлен поисковый запрос из нескольких словосочетаний, соответствующий тематике выбранного текста: «уменьшение объема текстовой информации», «обработка текста», «информационная система».

Полученные данные в ходе экспериментального исследования и их интерпретация

Пример оценок, сформированных системой для выбранного текста, приведен на рис. 5. В первой колонке выводится номер предложения. Оценка 1 обозначает соответствие предложения текста запросу пользователя, оценка 2 – степень повторяемости всех слов, оценка 3 – степень повторяемости существительных.

Графики, соответствующие трем полученным оценкам, представлены на рис. 6. Номера предложений на графиках соответствуют порядку следования предложений в тексте.

График на рис. 6 построен на основе таблицы, приведенной на рис. 5. Использованы оценки тех предложений, значения которых (в процентах) больше нуля. Из полученных результатов следует, что оценка, полученная по поисковому запросу пользователя, редко совпадает с оценками 2 и 3, основанными на статистике повторяемости слов в предложениях текста.

На рис. 7 представлен график, отображающий аналогичное сравнение оценок, однако в данном случае отображены все предложения текста, которые приведены в порядке их появления в тексте.

Для заданного поискового запроса объем полученного текста в результате сокращения информации составляет 22 % от объема исходного текста. Из графика следует, что основная доля предложений, соответствующих запросу пользователя, представлена в заключительной части текста, что позволяет исследователю обратить внимание именно на эту часть анализируемого документа.

№ (предл.)	Оценка 1 (%)	Оценка 2 (%)	Оценка 3 (%)	Предложение
3	100	25	31	Примерами видов обработки текста являются...
4	100	33	34	Для уменьшения затрат времени на обработк...
5	100	65	64	Таким образом, существует потребность в ин...
7	100	66	52	Основной целью разработки описываемой и...
11	100	24	16	В ходе работы были определены основные ф...
35	100	49	55	В качестве основы для построения модуля с...
67	100	100	83	В результате применения модуля поиска совп...
1	50	31	20	В качестве информационных источников мог...
0	50	27	19	С развитием информационных технологий в...
28	50	64	73	Данный текст по объему намного меньше ис...
56	50	48	52	Вкладка содержит исходный текст, выбранны...
34	50	54	61	Модуль сжатия представляет собой итерацио...
58	50	100	100	Вкладка отображает абзацы оригинального т...
60	50	57	40	5 представлен результат информационного а...
2	50	26	25	Возникает необходимость обработки пользов...

Рис. 5. Оценки предложений фрагмента текста

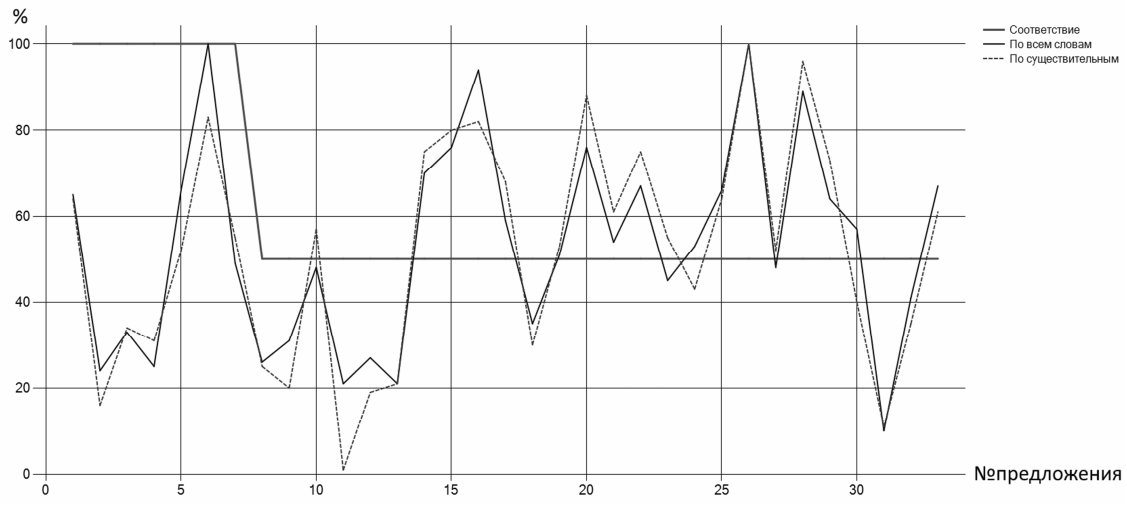

Рис. 6. Оценки предложений текста (по степени соответствия)

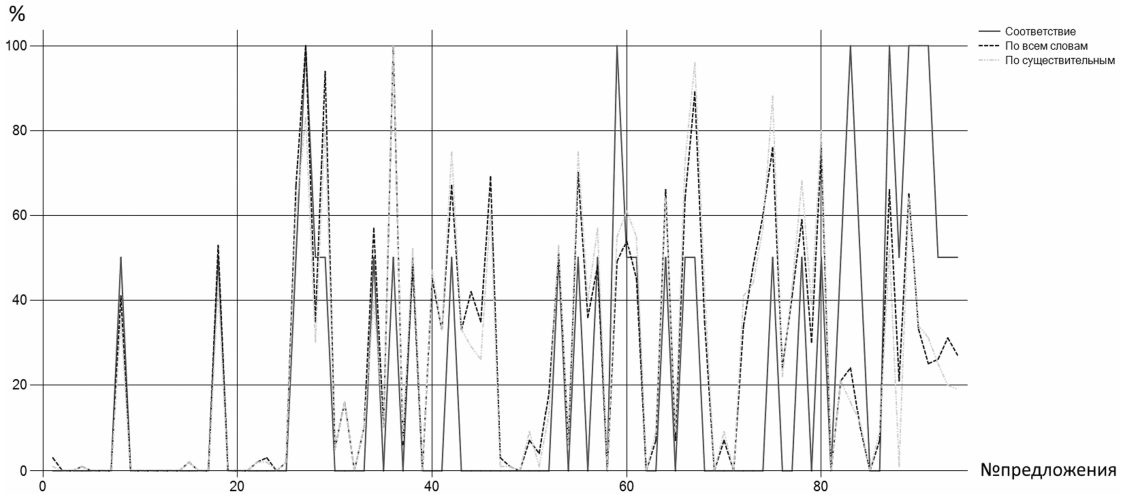

Рис. 7. Оценки предложений текста (по порядковым номерам)

Выводы

Таким образом, разработана информационная система, которая сочетает в себе несколько подходов к выделению полезной для пользователя информации и позволяет:

1. Сокращать объем анализируемого текста.
2. Выделять области текста, наиболее полно отвечающие запросу пользователя.
3. Получать результаты, которые могут быть использованы для составления рефератов и аннотаций анализируемых документов.

В дальнейшем предполагается формировать авторские смысловые группы слов, которые могут быть использованы исследователем для синтеза новых знаний.

Библиографические ссылки

1. *Алексеев А. А.* Тематическое представление новостного кластера как основа для автоматического аннотирования // Труды 15-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL)». 2013. С. 359–369.

2. *Бледнов А. М.*, *Моченов С. В.*, *Луговских Ю. А.* Об одном методе статистической фильтрации текстовой информации // Материалы междунар. науч. конф. «Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам (Ижевск, 13–17 июля 2006 г.). Ижевск : Изд-во ИжГТУ, 2006. С. 126–130.

3. *Герте Н. А.*, *Курушин Д. С.*, *Нестерова Н. М.* Моделирование понимания текста как основа автоматизированного реферирования // Материалы VII Междунар. науч. конф. «Индустрия перевода» (1–3 июня 2015 г.). Пермь : Изд-во Пермского нац. иссл. политех. ун-та, 2015. С. 81–84.

4. *Герте Н. А.* Методика денотативного анализа текста как возможный инструмент для автоматического реферирования // Вестник Российского нового университета. Серия «Человек в современном мире». 2015. Вып. 3. С. 35–38.

5. *Hong K. and Nenkova A.* Improving the Estimation of Word Importance for News Multi-Document Summarization // EACL. 2014. Pp. 712-721. URL: https://repository.upenn.edu/cgi/viewcontent.cgi?article=2036&context=cis_reports (дата обращения: 14.03.2018).

6. *Rankel P.*, *Dang H.*, *Conroy J.*, *Nenkova A.* A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art // 51st Annual Meeting of the Association for Computational Linguistics. 2013. Pp. 131-136. URL: http://newdesign.aclweb.org/anthology/P/P13/P13-2024.pdf (дата обращения: 14.03.2018).

7. *Luhn H. P.* The automatic creation of literature abstracts // IBM Journal of Research and Development. 1958. Vol. 2, no. 2, pp. 159-165. URL: https://text-analysis.googlecode.com/files/luhn58.pdf (дата обращения: 14.03.2018).

8. *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.

9. *Выдрин Д.*, *Громов С.*, *Поляков В.* Метод сравнения библиографических описаний, представленных в различных форматах // Обработка текста и когнитивные технологии № 9 : VII Междунар. конф. Варна ; М. : Учеба, 2004. С. 166–172.

10. *Выдрин Д.*, *Поляков В.* Реализация электронного словаря на основе н-грамм // Труды III Междунар. науч.-практ. конф. «Искусственный интеллект – 2002» / Ин-т проблем искусственного интеллекта, 2002. Т. 2. С. 79–84.

References

1. Alekseev A. A. (2013). Thematic representation of a news cluster as a basis for automatic annotation. Proceedings of the *Elektronnye biblioteki: perspektivnye metody I tekhnologii, elektronnye kollektsii*, pp. 359-369 (in Russ.).

2. Blednov A. M., Mochenov S. V., Lugovskikh Yu. A. (2006) About one method of statistical filtering of textual information. Proceedings of the *Sovremennye informatsionnye tekhnologii i pis'mennoe nasledie: ot drevnikh rukopisei k elektronnym tekstam*, pp. 126-130 (in Russ.).

3. Gerte N. A., Kurushin D. S., Nesterova N. M. (2015) Modeling the understanding of text as the basis for automated abstracting. Proceedings of the *Industriya perevoda*, pp. 81-84 (in Russ.).

4. Gerte N. A. (2015). Technique of denotative text analysis as a possible tool for automatic abstracting. *Vestnik Rossiiskogo novogo universiteta. Seriya: Chelovek v sovremennom mire* [Bulletin of the Russian New University. Series: Human in the modern world], no. 3, pp. 35-38 (in Russ.).

5. Hongand K., Nenkova A. (2014). Improving the Estimation of Word Importance for News Multi-Document Summarization in EACL, pp. 712-721, available at https://repository.upenn.edu/cgi/viewcontent.cgi?article=2036&context=cis_reports (accessed March 14, 2018).

6. Rankel P., Dang H., Conroy J., Nenkova A. (2013). 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, pp. 131-136, available at http://newdesign.aclweb.org/anthology/P/P13/P13-2024.pdf (accessed March 14, 2018).

7. Luhn H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*. Vol. 2, no. 2, pp. 159-165, available at https://text-analysis.googlecode.com/files/luhn58.pdf (accessed March 14, 2018).

8. Vtyurin M. V., Yastrebov A. I., Mochenov S. V. (2017). Development of an information system to reduce the volume of textual information in the process of information retrieval. *Intellektual'nye sistemy v proizvodstve* [Intelligent Systems in Manufacturing], vol. 15, no. 3, pp. 94-99 (in Russ.).

9. Vydrin D., Gromov S., Polyakov V. (2004). Method for comparing bibliographic descriptions presented in various formats. Proceedings of the *Obrabotka teksta i kognitivnye tekhnologi*i, pp. 166-172 (in Russ.).

10. Vydrin D., Polyakov V. (2002). The implementation of an electronic dictionary based on n-gram. Proceedings of the *Iskusstvennyi intellekt*, pp. 79-84 (in Russ.).

The Use of Statistical Characteristics to Reduce the Volume of Textual Information while Preserving its Informativeness

M. V. Vtyurin, Post-graduate, Kalashnikov ISTU, Izhevsk, Russia
S. V. Mochenov, PhD in Engineering, Kalashnikov ISTU, Izhevsk, Russia

The paper examines the possibility of researchers using specialized algorithms for an information system that provides a reduction in the volume of the analyzed text information in the process of information retrieval. The relevance of the work is justified by the complexity of information retrieval associated with the user's solution of a particular task and by the need to process large amounts of text data. The goal is to reduce the volume of the analyzed text information of Russian-language texts, while preserving their semantic component.

The description of the previously developed information system for reducing the volume of textual information in the process of information retrieval is given. A description of two different approaches to text analysis is presented, which allows for a comparative analysis of the results obtained. These approaches were implemented based on the previously developed information system. Corresponding changes were made in the structural scheme and algorithm of the information system functioning.

The results of the experimental study are presented. It follows from the results of the application of this approach that the main part of the proposals corresponding to the user's request for the selected text is shown in the final part of the text, which allows the researcher to pay attention to this part of the analyzed document. Results that can be used to compose abstracts and annotations of analyzed documents are obtained. In the future it is supposed to form author's semantic groups of words that can be used by the researcher to synthesize new knowledge.

Keywords: text analysis, information system, text information, reduction of the volume of the text, information search.

Получено 22.03.2018