УДК 004.912

Бызова А.К., Гольдштейн С.Л., Грицюк Е.М.

РАЗВИТИЕ ПОДСИСТЕМЫ СТРУКТУРИЗАЦИИ ТЕКСТОВ В СОСТАВЕ АВТОМАТИЗИРОВАННОГО ГЕНЕРАТОРА СИСТЕМНО ОБОСНОВАННОГО ТЕХНИЧЕСКОГО ЗАДАНИЯ

Аннотация. Рассматривается проблема обработки текстов, в частности, при составлении технических заданий. Проведен литературно-аналитический обзор возможных аналогов структуризации текстов. Разработаны системно-структурные и алгоритмические модели структуризации текстов с программной реализацией. Проведены испытания программного продукта.

Ключевые слова: структуризация, классификация, маркер.

Введение

Существует потребность обработки большого количества текстовой информации и оценки её релевантности для создания технических заданий (ТЗ), например, на информационные системы и программное обеспечение. Для поддержки этой работы создан автоматизированный генератор системно-обоснованного технического задания (АГ СО ТЗ), в структуру которого входит система электронизации входной информации (СЭВИ) [1, 2] с целью сбора, хранения и предварительной подготовки текстов.

Подсистемы СЭВИ АГ СО ТЗ (копирования, ввода с клавиатуры, сохранения, ранжирования) не дают возможности объединения данных из них в один документ и автоматической коррекции документов, переводимых в электронный вид. Поэтому в настоящей статье предлагается развить СЭВИ АГ СО ТЗ путем внедрения в её состав подсистемы структуризации текстов.

Литературно-аналитический обзор с выходом на аналоги и прототипы

Под структуризацией понимают компьютерную экспликацию лингво-полиграфического разбиения вербального текста [3]. Некоторые возможности структуризации текста заложены в программе MS Word. При литературно-аналитическом обзоре нами не найдено единого прототипа, позволяющего выполнить функции автоматической коррекции и объединения данных в один документ. Лучшие аналоги подсистем структуризации и каждой её составляющей (модулей и блоков) нами выбраны в качестве прототипов (таблица 1).

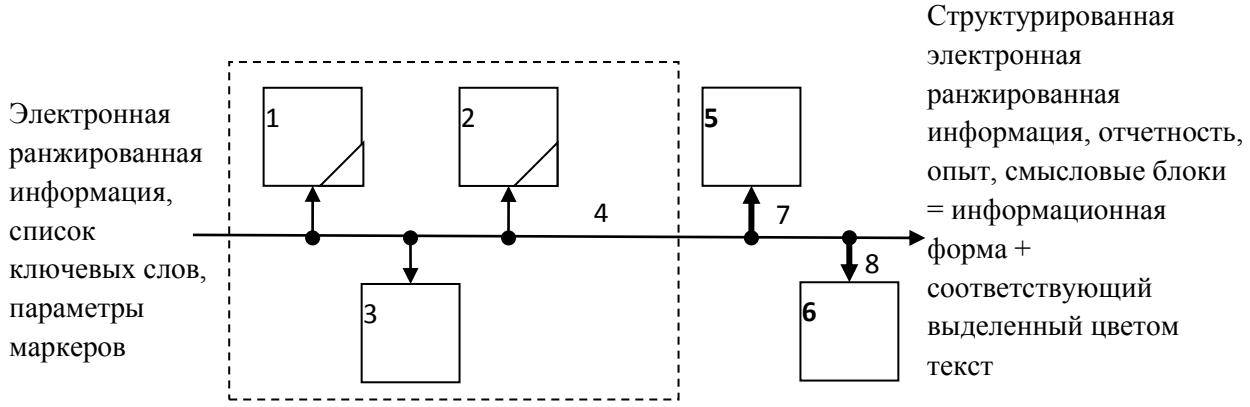
Предложения по парированию недостатков сформулированы нами далее на основе системно-структурных и алгоритмических моделей.

Таблица 1 – Пакет прототипов

Ранг прото-типа	Название подсистемы/ модуля	Название прототипа	Источник инфор-мации	Критика
0	Подсистема структуризации	Поисковая система FileSearchy	[4,5]	Общая критика: не дают возможность автоматизировано объединить файлы в один документ с одновременным поиском фрагментов текста по его смысловому содержимому с последующим редактированием
		Текстовый процессор MS Word	[6]	
		Способ оценки информации по Веревченко	[7]	
		Программа по объединению документов FileMerger	[8]	
1	Модуль параметров поиска	Текстовый процессор MS Word	[6]	Отсутствует маркировка результатов поиска (смысловых блоков) единым фрагментом
	Модуль редактирования текста			Отсутствуют группировка и выборка по маркерам нужных по смыслу фрагментов текста

Системно-структурные модели

На основе рассмотренных аналогов нами составлен компилятивный прототип подсистемы структуризации и предлагается его улучшение (Рис. 2).



Модули: 1 – выбора файла; 2 – поиска смысловых блоков; 3 – оценки; 5 – ввода параметров поиска; 6 – редактирования; 4, 7, 8 – интерфейсов.

Рисунок 2 – Системно-структурная модель подсистемы структуризации по прототипу и предлагаемому решению (штриховка)

Усовершенствованная подсистема структуризации дает возможность автоматизированного объединения релевантно-пертинентных фрагментов текста и их редактирования.

Алгоритмическая модель

На Рисунке 3 представлена алгоритмическая модель подсистемы структуризации на языке блок-схем. Алгоритм подсистемы структуризации организован циклически по маркерам (5 и 19) и подразумевает последовательный вызов 5 основных процедур: 3 – выбор файла, 7 – задание параметров поиска необходимых фрагментов текста, 9 – непосредственно поиска с выделением цветом результатов поиска, 13 – редактирование маркеров и текста, 15 – оценка полноты и качества полученной информации. При этом под маркером будем понимать информационную форму, атрибутами которой считаем цвет, название и ключевое слово. Найденные по ключевому слову маркера фрагменты текста будут выделены согласно цвету маркера.

Компьютерная реализация

На основе алгоритма спроектирована дополнительная подсистема структуризации для АГ СО ТЗ и осуществлена её программная реализация в среде Microsoft Visual Studio 2013 на языке C#.

Для начала работы необходимо выбрать в выпадающем списке «Электронизация входной информации» пункт «Структуризация». Тогда в новом окне откроется экранная форма подсистемы структуризации.

Для того чтобы объединить несколько документов в один, в окне «Выбор нужной папки» выбираем папку, которая содержит файлы для объединения. Также можно воспользоваться поиском файлов по ключевому слову в окне «Поиск документов». Содержимое выбранной папки либо результаты поиска будут выведены в соответствующем окне. При нажатии на кнопку «Добавить» выбранные файлы будут выведены в окне «Файлы для объединения». При нажатии на кнопку «Удалить» отмеченные галкой файлы для объединения будут удалены из списка файлов для объединения. При нажатии на кнопку «Объединить» выделенные галкой файлы будут объединены в новом файле. В диалоговом окне будет предложено место сохранения этого нового файла.

В окне «Параметры маркеров» представлена таблица маркеров. Каждой строке соответствует маркер, для которого выбирается свой собственный цвет выделения, название, ключевое слово для автопоиска нужных фрагментов текста при нажатии на кнопку «Выделить». Предусматривается возможность ручной маркировки с помощью кнопки «Вручную». Можно добавлять новые маркеры и удалять, и копировать маркеры, отмеченные галкой.

Рисунок – 3 Алгоритмическая модель подсистемы структуризации по прототипу и предлагаемому решению (штриховка)

При группировке объединяются фрагменты текста, выделенные одним цветом и расположенные в разных частях документа. В окне «Редактирование» в начале документа будут показаны фрагменты, выделенные одним цветом, объединенные в более крупные блоки, текст без маркировки располагается в конце документа. При нажатии «Показать выбранное» в окне «Редактирование» будут отображаться только те маркированные фрагменты текста, которые перед этим отметили в таблице, расположенные в документе в том же порядке, что и перед редактированием.

Вычислительный эксперимент

Дано: имеется 2 текста: концепция профилактики инфекций, связанных с оказанием медицинской помощи (концепция проф исмп.txt) и положение о враче-эпидемиологе амбулаторно-поликлинического учреждения (93_09_17_N220_11.txt). Скорость чтения в среднем у взрослого человека примем 180 слов в минуту [9].

Требуется: 1) определить время T1, которое необходимо специалисту для прочтения исходных текстов; 2) определить время T2, которое потребуется специалисту для прочтения исходных текстов с использованием подсистемы структуризации; 3) сравнить T1 и T2.

Рассмотрим работу в подсистеме структуризации.

Объединили два исходных текста, далее работаем с полученным объединенным текстом (merge_text.txt). Задали три маркера для трех ключевых слов: «дезинфекция», «иммунизация», «стерилизация». Для учета словооснов окончания ключевых слов отбросили, т.е. «дезинфекци», «иммунизаци», «стерилизаци». Провели автоматический поиск соответствующих фрагментов текста для каждого ключевого слова. В результате получили структурированный текст (struct_text.txt), состоящий из четырех блоков: три смысловых блока, соответствующие маркерам, и один немаркированный – остаток исходного текста.

Результаты эксперимента приведены в таблице 2.

Таблица 2 – Характеристики результатов

Исследуемые тексты	Значения характеристик	
	Количества слов	Времени необходимого для прочтения, (мин)
концепция проф исмп.txt	6435	36
93_09_17_N220_11.txt	564	3
merge_text.txt	6999	39
struct_text.txt	1191	7

Время необходимое для прочтения обоих текстов будет равно сумме времен, необходимых для прочтения каждого текста в отдельности, или времени для прочтения объединенного текста, т.е. T1 = 36 + 3 = 39. Время для прочтения релевантно-пертинентных фрагментов текста T2 = 7. Видно, что специалист тратит в 5,14 раз меньше времени на отобранных компьютером фрагментов исходного текста при использовании подсистемы структуризации. Таким образом, функции объединения данных из них в один документ, поиска фрагментов текстов по ключевым словам, и возможность их последующего редактирования экономят служебное время специалиста.

Пример использования программы представлен на Рисунке 4.

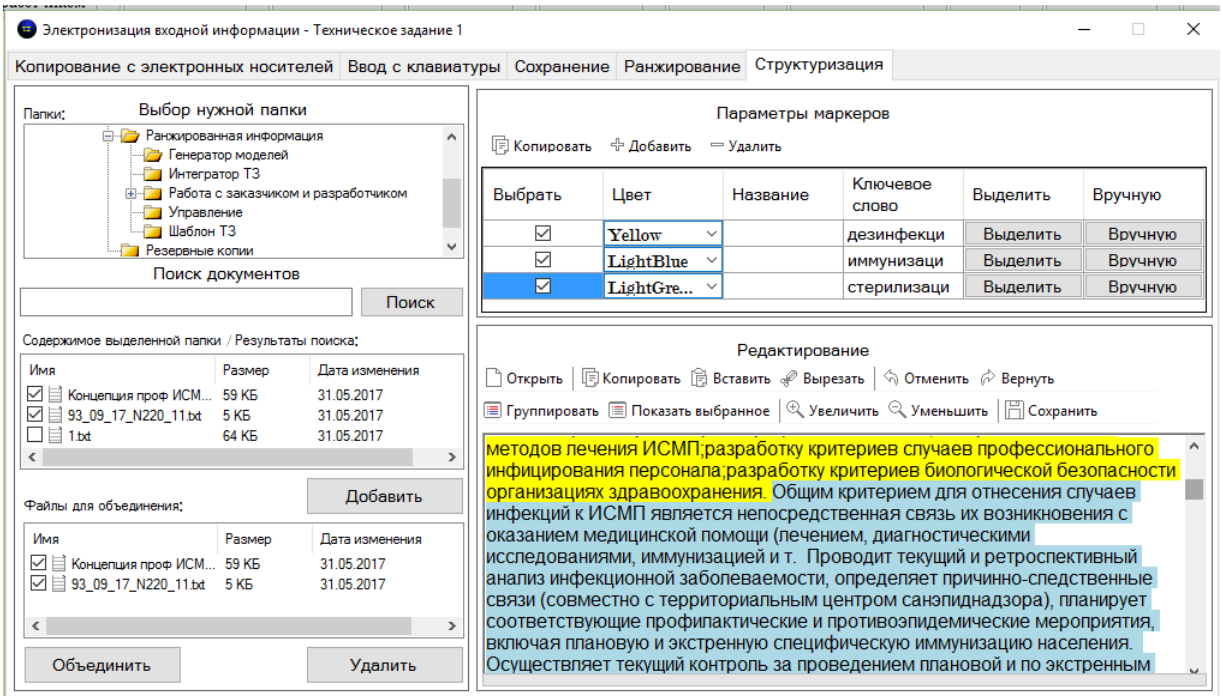


Рисунок 4 – Экранная форма подсистемы структуризации

Заключение

В ходе работы был:

- проведен литературно-аналитический обзор, в результате которого найдены аналоги подсистемы, составлен компилятивный прототип и проведена его критика;
- предложен пакет системно-структурных и алгоритмических моделей подсистемы структуризации;
- на основе пакета моделей разработан программный продукт;
- проведен вычислительный эксперимент, результаты которого показали, что с использованием подсистемы структуризации специалист тратит в 5,14 раз меньше времени на прочтение релевантно-пертинентных фрагментов текста.

Таким образом, спроектированная подсистема структуризации текстов предоставляет возможности автоматизированного объединения файлов в один документ, поиска фрагментов текста по его смысловому содержимому с последующим редактированием, маркировки результатов поиска (смысловых блоков) единым фрагментом, группировки и выборки по маркерам нужных по смыслу фрагментов текста.

Библиографический список

1. Гольдштейн С. Л. Развитие системы электронизации входной информации [Электронный режим] / С. Л. Гольдштейн, Е. М. Грицюк, Д. А. Леонов // Системная интеграция в здравоохранении : электрон. науч. журн. – 2012. – № 2. – С. 5–18. – Режим доступа: http://www.sys-int.ru/files/2012.2/152/sys_int_136_2_12_2012.pdf.
2. Подсистема электронизации входной информации в автоматизированном генераторе системно-обоснованного технического задания : свидетельство о гос. регистрации программы для ЭВМ № 2013612950 / Гольдштейн С. Л., Грицюк Е. М., Леонов Д. А. ; правообладатель Урал. федер. у-т им. первого Президента России Б. Н. Ельцина.– № 2013610618 ; заявл. 01.02.2013 ; регистрация 19.03.2013 ; опубл. 20.06.2013.
3. Гринбаум О. Н. Структуризация текста в компьютерной системе «ЛИНДА» / О. Н. Гринбаум, Г. Я. Мартыненко // Структурная и прикладная лингвистика : межвуз. сб. / под ред. А. С. Герда. – Санкт-Петербург : Изд-во С.-Петерб. ун-та, 1993. – Вып. 4. – С. 171–181.
4. FileSearchy 1.43 [Электронный ресурс] : программа для поиска файлов на компьютере // SoftPortal : офиц. сайт. – Режим доступа: http://www.softportal.com/software-33494-filesearchy.html.
5. FileSearchy.com [Электронный ресурс]. – Режим доступа: http://www.filesearchy.com/ru/.
6. Microsoft Word [Электронный ресурс] // Википедия : свободная энцикл. – Режим доступа: https://ru.wikipedia.org/wiki/Microsoft_Word.
7. Веревченко А. П. Информационные ресурсы для принятия решений / А. П. Веревченко [и др.]. – Москва : Академ. проспект, 2002. – 560 с.
8. Быстро объединяем файлы Word в один документ [Электронный ресурс] // MoyTop.com : блог интернет специалиста. – 2015. – Режим доступа: http://moytop.com/rerajting-i-kopirajting/1-5/bystro-obedinyaem-fajly-word-v-odin-dokument.
9. Душков Б. А. Быстрое чтение / Б. А. Душков // Психология труда, управления, инженерная психология и эргономика : энцикл. слов. / Б. А. Душков, А. В. Королев, Б. А. Смирнов. – Москва : Академ. проект ; Деловая кн., 2005. – С. 183–185.