УДК 004.912

# О СНИЖЕНИИ ЗАТРАТ ВРЕМЕНИ НА АНАЛИЗ И ОБРАБОТКУ ТЕКСТА СПЕЦИАЛИСТОМ С ИСПОЛЬЗОВАНИЕМ КОМПЬЮТЕРНОЙ ПОДДЕРЖКИ

**Грицюк Е.М.[1], Гольдштейн С.Л.[2], Дугина Е.А.[1], Бызова А.К.[2]**

*[1] ГАУЗ СО МКМЦ «Бонум», г. Екатеринбург, Россия*
*[2] ФГАОУ ВПО «УрФУ», г. Екатеринбург, Россия*

Рассмотрено решение задачи снижения затрат времени специалистом на анализ и обработку текста при использовании средств компьютерной поддержки. Сформулированы условия задачи и проведен эксперимент. Представлены структура и алгоритм предлагаемого решения. Получены и оценены результаты сравнения двух способов анализа и обработки текстов.
**Ключевые слова**: анализ и обработка текстов, компьютерная поддержка анализа текстов, поиск и структуризация данных, затраты времени, моделирование.

## About reducing time spent on analysis and processing of text by a specialist using computer support

**Gritsyuk E.M.[1], Goldshtein S.L.[2], Dugina E.A. [1], Byzova A.K.[2]**

*[1] State Autonomic Health Institution of the Sverdlovsk Region Multiprofile Clinical Medical Center «BONUM»*
*[2] Ural Federal University, Yekaterinburg, Russia*

The solution to the problem of reducing the time to analysis and text processing when using the means of computer support. Formulated the conditions of the problem and the experiment. Presents the structure and algorithm of the proposed solution. Obtained and reviewed results of the comparison of two methods of analysis and processing of texts.
**Keywords:** analysis and word processing, computer assisted text analysis, search and structuring of data, time, modeling.

### Введение

Анализ и обработка текстовых документов (особенно большого объема) – актуальная задача для многих специалистов из разных сфер деятельности, в том числе, в медицине. Здесь она может возникать особенно часто перед врачами-организаторами здравоохранения и (или) сотрудниками, занимающимися научно-исследовательской деятельностью. При этом, как правило, помимо работы с текстами имеет место выполнение основных функций и обязанностей, т.е. временной ресурс ограничен. Поэтому актуально снизить затраты времени на анализ и обработку объемных текстовых документов (по возможности за счет средств компьютерной поддержки) для более рационального использования временного ресурса: например, потратить на

ранжирование обработанных материалов с целью дальнейшего хранения, преобразования их в другой документ, формирования некоего управленческого решения и др.

**Формулировка задачи**

Представляя условия задачи обработки и анализа текстов в упрощенном виде, получим следующее:

дано: специалист и его компетенции (согласно образовательным и профессиональным стандартам), а также другие ресурсы (в том числе временной) на решение задачи, текстовый документ (произвольный по объему, форме и содержанию), пакет запросов на анализ и обработку текста, критерии оценки качества решения; инструмент – автоматизированный (компьютерный) и неавтоматизированный (способ «вручную»);

требуется: оценить и сравнить затраты времени на анализ и обработку текстового документа автоматизированным и неавтоматизированным способами.

**О критериях оценки качества**

В соответствии с поставленной задачей основной критерий качества анализа и обработки текстовых документов – сокращение времени. Его элементы предлагается рассматривать исходя из формализма SADT (IDEF 0)– технологий [1] в виде кортежных моделей:

$$ТСВ = <X_1, X_2, X_3, X_4, X_5; R>, \quad (1)$$

где ТСВ – технология сокращения затрат времени при анализе и обработке текстовых документов за счет резервов ее составляющих: $X_1$ – на сырьевом входе, $X_2$ – на управленческом входе, $X_3$ – на исполнительном входе, $X_4$ – в основном процессе и $X_5$ – в выходе продукта;

$$X_1 = <X_{11}, X_{12}, X_{13}; R_1>, \quad (2)$$

где резервы времени за счет: X11 – разделения исходного текстового документа на части по числу задействованных специалистов, X12 – формализации пакета запросов, X13 – структуризации тезауруса специалиста;

$$X_2 = <X_{21}, X_{22}, X_{23}; R_2>, \quad (3)$$

где резервы времени за счет формализации: X21 – нормативов, X22 – методик анализа и обработки документов, X23 –иерархии руководителей;

$$X_3 = <X_{31}, X_{32}, X_{33}; R_3>, \quad (4)$$

где резервы времени за счет: X31 – привлечения специалиста с дополнительными компетенциями аналитика, X32 – оборудованного персонального рабочего места (в плане

hard/soft - составляющих) адекватно поставленной задаче, X33 – рациональной иерархии исполнителей;

$$X_4 = <X_{41}, X_{42}, X_{43}, X_{44}, X_{45}; R_4>, \quad (5)$$

где резервы времени за счет автоматизации: X41 – поиска и импорта документов, X42 – объединения нескольких документов в один, X43 – маркировки текста в соответствии с контентом, X44 – группировки маркированных фрагментов текста с похожим по смыслу контентом, X45 – выборки маркированных фрагментов текста с разным по смыслу контентом;

$$X_5 = <X_{51}, X_{52}, X_{53}, R_5>, \quad (6)$$

где резервы времени за счет текстовых шаблонов: X51 – «сырья», т.е. структуры анализируемого документа, X52 – плана работы (алгоритма мероприятий), X53 – «готовых продуктов», т.е. различных директивных документов (писем, приказов и др.) либо других форм (статей, диссертаций и др.), $R_{1-5}$ – матрицы связей.

При формулировке пакета запросов необходимо учитывать глубину анализа, актуальную сложившейся ситуации с лимитом времени. В свою очередь глубина анализа зависит от сложности иерархии понятий, отраженных в моделях деятельности специалиста (за счет числа уровней, вершин и связей [2]), пример приведен на рис.1.

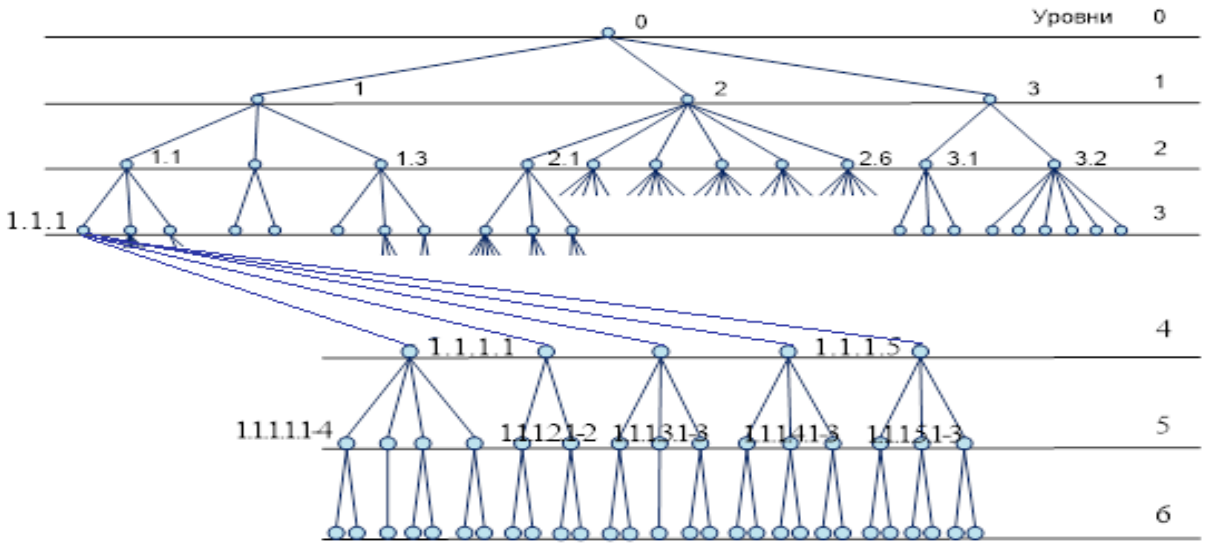


Рис.1 Иерархия понятий госпитального эпидемиолога (фрагмент) [3]

(0-система противоэпидемической поддержки (СПЭП); 1-объекты деятельности эпидемиолога (ресурсы), 2-многоролевая деятельность госпитального эпидемиолога (МДГЭ), 3-оценка деятельности; 1.1-материальные, 1.2-человеческие, 1.3-информационные; 1.1.1-оборудование, 1.1.2-материалы, 1.1.3-здание; 1.2.1-сотрудники, 1.2.2-пациенты; 1.3.1-информационная база, 1.3.2-компьютерное обеспечение, 1.3.3.-директивные документы; 2.1-в роли врача, 2.2-в роли педагога, 2.3-в роли менеджера, 2.4-в роли научного работника, 2.5-в роли проектировщика, 2.6-в роли системного интегратора; 2.1.1-информационная функция, 2.1.2-диагностическая функция, 2.1.3-управленческая функция; 3.1-оценка процесса, 3.2-оценка результата; 3.1.1-оценка технологичности процесса, 3.1.2-оценка затратности процесса, 3.1.3-оценка своевременности процесса; оценки результата деятельности эпидемиолога: 3.2.1-оценка результата деятельности врача, 3.2.2-оценка результата деятельности педагога, 3.2.3-оценка результата деятельности менеджера, 3.2.4-оценка результата деятельности научного работника, 3.2.5-оценка результата деятельности проектировщика, 3.2.6-оценка результата деятельности системного интегратора); 1.1.1-оборудование, 1.1.1.1-оборудование дезинфекционное, 1.1.1.2-оборудование стерилизационное, 1.1.1.3-оборудование вспомогательное для обеззараживания, 1.1.1.4-оборудование лечебно-диагностическое, 1.1.1.5-оборудование для утилизации отходов класса Б и В; 1.1.1.1.1-оборудование дезинфекционное для воздуха, 1.1.1.1.2-оборудование дезинфекционное для поверхностей, 1.1.1.1.3-оборудование дезинфекционное для инструментов, 1.1.1.1.4-оборудование дезинфекционное для постельных принадлежностей; 1.1.1.1.1.1-оборудование дезинфекционное для воздуха открытого типа, 1.1.1.1.1.2-оборудование дезинфекционное для воздуха закрытого типа; 1.1.1.1.2.1-оборудование дезинфекционное для поверхностей аэрозольное; 1.1.1.1.3.1-оборудование дезинфекционное для инструментов ультразвуковое, 1.1.1.1.3.2-оборудование дезинфекционное для инструментов автоматическое; 1.1.1.1.4.1 оборудование дезинфекционное для постельных принадлежностей пароформалиновое, 1.1.1.1.4.2-оборудование дезинфекционное для постельных принадлежностей озоновое; 1.1.1.2.1-оборудование стерилизационное высокотемпературное, 1.1.1.2.2-оборудование стерилизационное низкотемпературное;1.1.1.2.1.1-суховоздушные стерилизаторы, 1.1.1.2.1.1-паровые стерилизаторы; 1.1.1.2.2.1-газовые стерилизаторы, 1.1.1.2.2.2-плазменные стерилизаторы; 1.1.1.3.1-оборудование обеззараживания вспомогательное для хранения стерильного, 1.1.1.3.2-оборудование обеззараживания вспомогательное для упаковки на стерилизацию, 1.1.1.3.3 прочее оборудование для дезинфекции и стерилизации; 1.1.1.3.1.1-оборудование обеззараживания вспомогательное для хранения стерильного ультрафиолетовое, 1.1.1.3.1.1-оборудование обеззараживания вспомогательное для хранения стерильного озоновое; 1.1.1.3.2.1-оборудование для упаковки на стерилизацию путем ламинирования; 1.1.1.3.3.1-контейнеры для дезинфекции, 1.1.1.3.3.2-коробки для стерилизации (биксы); 1.1.1.4.1-оборудование для инвазивных манипуляций, 1.1.1.4.2-оборудование для анестезиологии и реанимации, 1.1.1.4.3-оборудование поверхностно контактирующее с кожей; 1.1.1.4.1.1-оборудование контактирующее со стерильными органами и тканями, 1.1.1.4.1.2-оборудование контактирующие с условно заселенными слизистыми поверхностями, 1.1.1.4.2.1-аппаратура для жизнеобеспечения, 1.1.1.4.2.1-кювезы для ухода за недоношенными детьми; 1.1.1.4.3.1-оборудование для физиотерапии, 1.1.1.4.3.1-оборудование для не инвазивного обследования (тонометры, термометры и др.); 1.1.1.5.1-оборудование для высокотемпературных методов утилизации отходов класса Б и В, 1.1.1.5.2-оборудование для других методов утилизации отходов класса Б и В, 1.1.1.5.3-вспомогательное оборудование для утилизации отходов класса Б и В; 1.1.1.5.1.1-инсинераторы, 1.1.1.5.1.2-деструкторы игл; 1.1.1.5.2.1-оборудование для утилизации отходов класса Б и В путем измельчения, уплотнения и обеззараживания, 1.1.1.5.2.2-оборудование для утилизации отходов класса Б и В с помощью СВЧ; 1.1.1.5.3.1-многоразовые контейнеры для отходов, 1.1.1.5.3.2-стойки-тележки для отходов

Для анализа и обработки текста подобная иерархия понятий в соответствии с профилем деятельности специалиста необходима. Она позволит существенно сэкономить время при формировании пакета запросов и ориентировке в наборе полученных ответов. Формирование иерархических моделей происходит при получении базового образования, затем они дополняются/уточняются опытом и системой непрерывного совершенствования знаний, в том числе и при научно-исследовательской работе. Храниться иерархии могут на бумажных или электронных носителях, а также в ментальном виде и использоваться практически. При этом качество анализа и обработки текстовых документов существенно зависит от сложности иерархических моделей понятий деятельности специалиста (табл. 1).

Таблица 1
Зависимость качества анализа от структурной сложности иерархии понятий деятельности специалиста [4]

| Сложность иерархии понятий | Качество анализа по | |
|---|---|---|
| | глубине | полноте |
| 1-2 уровня, число вершин 7 ± 2, связей до 72 (максимально возможное количество) | поверхностный | скрининговый |
| 1-4 уровней, число вершин до 800, связей до 640 000 (максимально возможное количество) | средний | средний |
| 1-6 уровней, число вершин до 66 300, связей до 4 300 тыс. | глубокий | полный |

Анализ информации, различный по глубине и полноте, может осуществляться либо последовательно по этапам (вначале – поверхностный/скрининговый, затем – средний и в конце – глубокий/полный), либо – реализуется только частично в соответствии с поставленной задачей. Например, когда необходимо провести обзор документов (вновь вводимых нормативных актов или научных публикаций по определенной тематике) с целью определения актуальности этих текстов для специалиста (врача-организатора или врача-исследователя) можно ограничиться поверхностным скринингом. Если нужно получить экспертное решение, составить техническое задание или получить какую-то другую форму экстрагированного знания (из одного или нескольких информационных источников), тогда надо проводить глубокий полный анализ текстов. Если ресурса времени у специалиста недостаточно, то выполнение задачи анализа и обработки текста затрудняется. При этом на основе модели понятий деятельности специалиста можно сформировать соответствующий ситуации пакет запросов, например, в виде когнитивной карты (рис. 2).

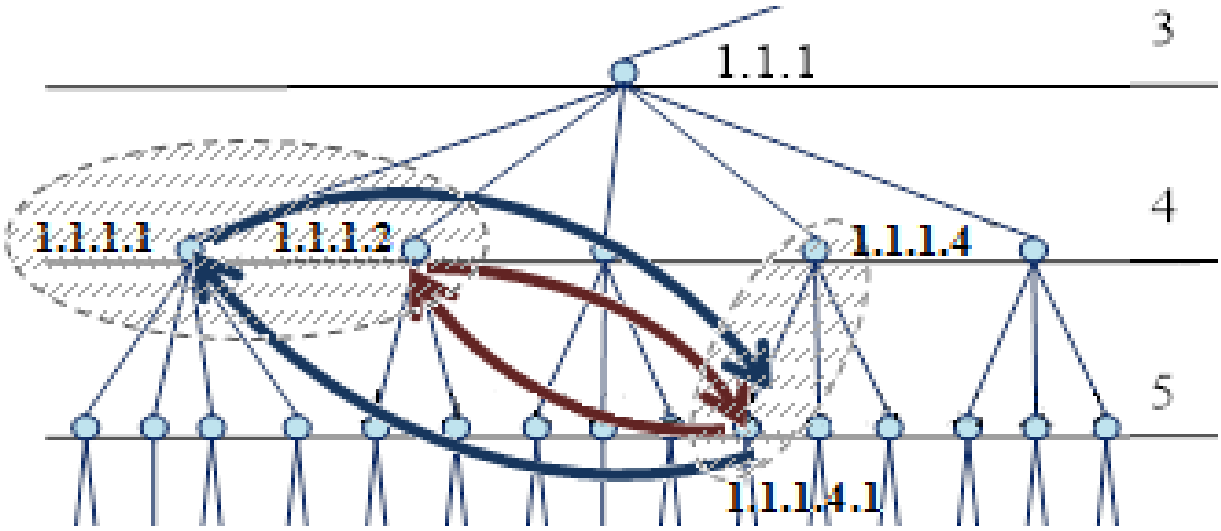


Рис. 2 Когнитивная карта запросов на информацию по оборудованию, участвующему в обеспечении инфекционной безопасности (штриховкой и пунктиром выделены интересующие области карты)

(1.1.1-оборудование, 1.1.1.1-оборудование дезинфекционное, 1.1.1.2-оборудование стерилизационное, 1.1.1.4-оборудование лечебно-диагностическое, 1.1.1.4.1-оборудование для инвазивных манипуляций)

Из когнитивной карты запросов на информацию по оборудованию, участвующему в обеспечении инфекционной безопасности, видно, что оборудование для дезинфекции (1.1.1.1) и стерилизации (1.1.1.2) имеет связь с лечебно-диагностическим оборудованием (например, для инвазивных манипуляций – 1.1.1.4.1). С одной стороны, выбор способов дезинфекции и стерилизации зависит от вида оборудования, которое должно таким образом обеззараживаться. С другой стороны, бывает и обратная связь: когда оборудование для лечения и диагностики выбирают в зависимости от имеющихся методов дезинфекции и стерилизации, обеспечивающихся специальным оборудованием. Обычно решающим фактором при выборе становится стоимость и целесообразность использования разного вида устройств. Как правило, не приобретают дорогостоящий плазменный стерилизатор только для обработки металлических шпателей в ЛОР-кабинете. Или обратная ситуация: при наличии суховоздушного стерилизатора целесообразно выбирать термостойкие оборудование/ инструменты для осуществления определенных медицинских технологий.

Плюсы построения когнитивной карты в том, что таким образом четко позиционируются и ограничиваются пределы при формирования запросов (не надо для каждой новой поставленной задачи строить и изучать всю объемную иерархию, например, по оборудованию как на рис. 1, а только ее небольшую интересующую специалиста часть как на рис. 2), а также повышается точность их формулировки.

## Моделирование компьютерной поддержки

Требования к алгоритму программного обеспечения взяты из опыта практической деятельности по обработке и анализу текстов «вручную»: чтение, выделение структурных единиц текста по запросу специалиста, структуризация релевантных единиц (фрагментов), маркировка фрагментов в соответствии с семантическим содержанием.

При литературно-аналитическом обзоре нами не найдено единого прототипа, позволяющего выполнить эти функции. Поэтому на основе рассмотренных аналогов [5-8] составлен компилятивный прототип и предложено его развитие (рис. 3).

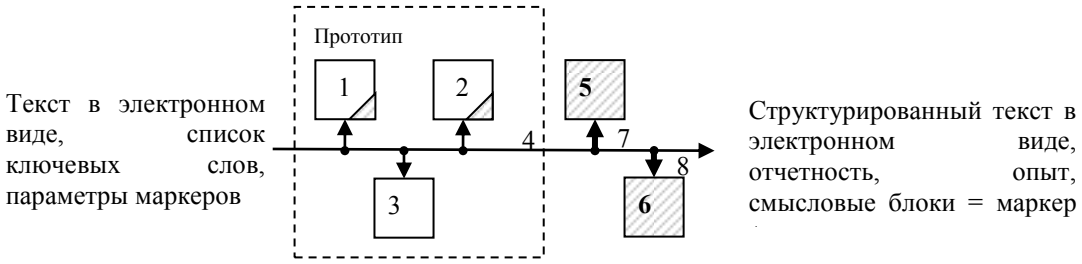


Рис. 3. – Системно-структурная модель прототипа и предлагаемого решения – выделено штриховкой (модули: 1 – выбора файла; 2 – поиска смысловых блоков; 3 – оценки; 5 – параметров специального поиска; 6 – редактирования; 4, 7, 8 – интерфейсов) [9]

Алгоритм функционирования (рис. 4) организован циклически по маркерам (блоки 5 и 19) и подразумевает последовательный вызов пяти основных процедур: блок 3 – выбора текстового файла для обработки, блок 7 – задания параметров поиска необходимых фрагментов текста, блок 9 – непосредственно поиска, блок 13 – редактирования маркеров и текста, блок 15 – оценки полноты и качества полученной информации [9]. Программная реализация выполнена нами в среде Microsoft Visual Studio 2013 на языке C#.

## Эксперимент оценки качества обработки и анализа текста

В основу эксперимента положена следующая служебная ситуация: врач-организатор составляет аналитический отчет, в который желательно ввести несколько фрагментов из двух нормативных документов: 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) [10]. Для эксперимента по обработке и анализу текста с использованием разных способов (не автоматизированного, т.е. «вручную» и автоматизированного, т.е. при помощи компьютерной поддержки) формулируем условия его проведения.

Дано: врач-организатор первой квалификационной категории с опытом работы более 10 лет, имеющий опыт аналитической работы с текстовыми документами, пакет моделей понятий по профилю деятельности (6-ти уровней детализации), необходимые ресурсы на выполнение задачи (оборудованное рабочее место, время и др.), пакет запросов на анализ и обработку документа, критерий оценки качества выполнения, два текстовых документа по профилю деятельности [10, 11] в электронном виде (6 435 знаков в первом и 564 знаков во втором) для анализа и обработки;

инструмент: автоматизированный (компьютерный) и неавтоматизированный (способ «вручную»);

Требуется: 1) определить время T1, которое необходимо специалисту для прочтения, анализа и обработки релевантно-пертинентных фрагментов текстов неавтоматизированным способом «вручную»; 2) определить время T2, которое потребуется специалисту для прочтения, анализа и обработки релевантно-пертинентных фрагментов исходных текстов автоматизированным способом с использованием компьютерной поддержки; 3) сравнить T1 и T2. Учесть, что скорость чтения в среднем у взрослого человека примерно 180 слов в минуту [11].

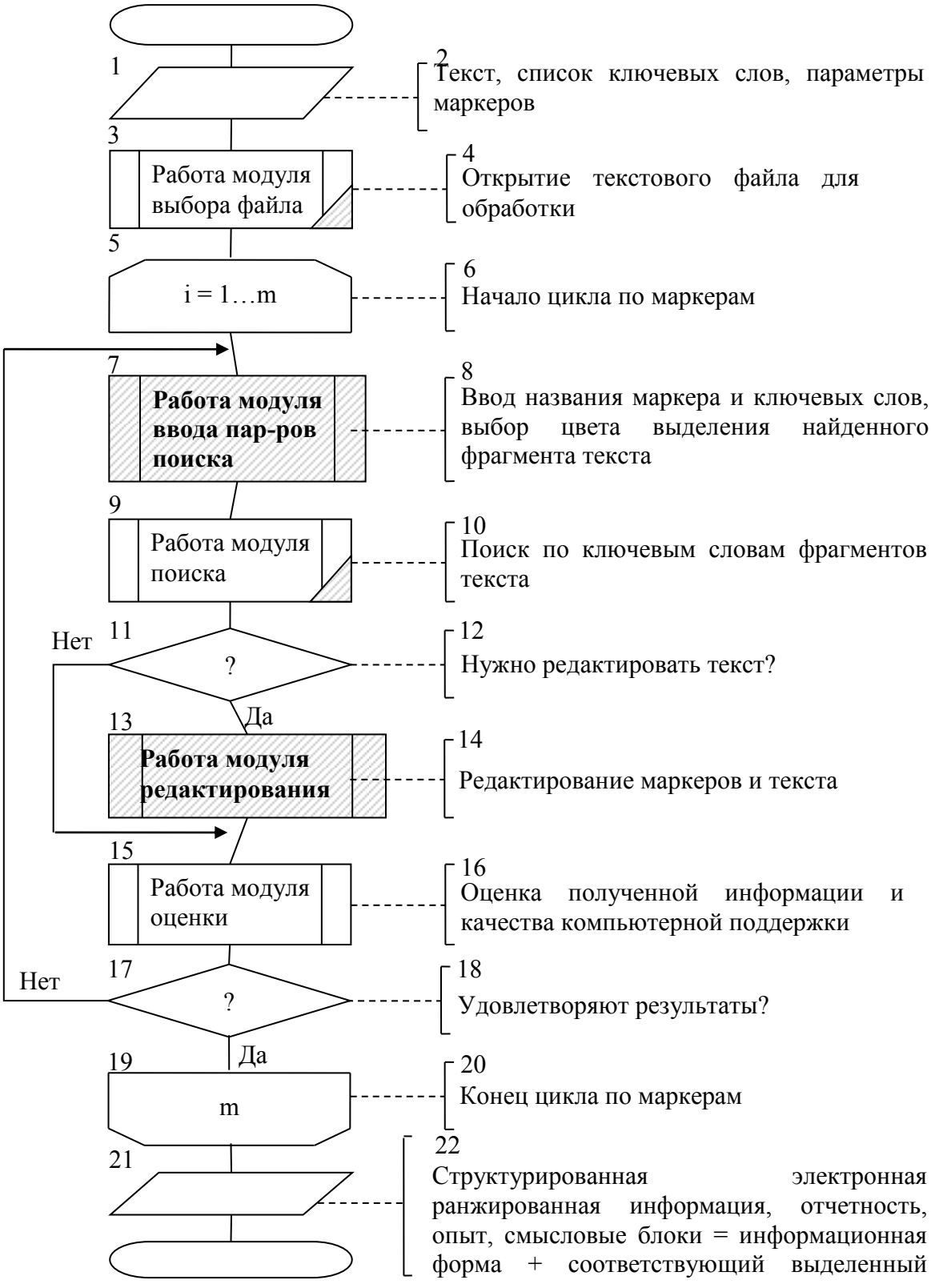


Рис. 4 – Алгоритмическая модель прототипа и предлагаемого решения (штриховка)

Результаты эксперимента приведены в таблице.

Таблица 2
Результаты эксперимента

| Исследуемые тексты | Значения характеристик: | | | | |
|---|---|---|---|---|---|
| | количество слов | время, необходимое для анализа и обработки, (мин) | | | |
| | | не автоматизированным способом «вручную» (T1) | | автоматизированным способом (T2) | |
| | | чтение | анализ и обработка | чтение | анализ и обработка |
| концепция проф исмп.txt | 6435 | 36 | 60 | 6 | 32 |
| 93_09_17_N220_11.txt | 564 | 3 | 8 | 1 | 4 |
| В сумме | 69990 | 107 | | 43 | |

Время, необходимое для прочтения, анализа и обработки обоих текстов, равно сумме времен для изучения каждого текста в отдельности или времени для прочтения, анализа и обработки объединенного текста, т.е. T1 = 36 + 3 + 60 + 8 = 107. Время для прочтения релевантно-пертинентных фрагментов текста T2 = 6 + 1 + 32 + 4 = 43. Видно, что специалист тратит примерно в 2,5 раза меньше времени на прочтение, анализ и обработку релевантно-пертинентных фрагментов текста при использовании компьютерной поддержки. При этом внимание специалиста не тратится на рутинную работу, которую выполняет программное средство, а может быть направлено на более тщательный анализ изучаемого документа.

Таким образом, можно заключить что время, затрачиваемое для анализа и обработки входящего документа, зависит от количественно-качественных характеристик текста, числа ключевых слов и соответствующих им маркеров, а также способа обработки (при автоматизированном – затраты времени значительно ниже).

**Результаты**

1) Поставлена и решена задача оценки и сравнения автоматизированного и неавтоматизированного способов анализа и обработки текстового документа по затратам времени.
2) Предложены критерии оценки, системно-структурная и алгоритмическая модели автоматизированного способа анализа и обработки текстового документа по прототипу и предлагаемому решению.
3) Рассмотрена конкретная ситуация, по данным которой проведен компьютерный эксперимент.

**Вывод**

Решение задачи достигнуто. Предлагаемые улучшения предоставляют возможность автоматизированного объединения файлов в один документ с одновременным поиском фрагментов текста по его сегментам с последующим редактированием, маркировкой

результатов поиска (смысловых блоков) единым фрагментом, группировкой и выборкой по маркерам нужных по смыслу фрагментов текста. При помощи эксперимента показано, что предложенная компьютерная поддержка экономит ресурсы профильного специалиста, в частности, из медицинской сферы, при работе с новыми объемными документами.

Список литературы

1. Методология функционального моделирования IDEFØ. Руководящий документ, - М: Росстандарт России, - 2000, - 67 с.

2. Саати Т. Порядок расчета показателей важности по методике анализа иерархий // Мир Знаний [Офиц. сайт]. URL: http://mirznanii.com/a/169939/metod-analiza-ierarkhiy-t-saati

3. Грицюк Е. М. Развитие многоролевой деятельности госпитального эпидемиолога путем ее моделирования / Медицина и здравоохранение: материалы III Международной научной конференции. – Казань: Бук, 2015. – С. 69-75. URL: http://www.moluch.ru/conf/med/archive/154/8118/

4. Ландэ Д.В. Поиск знаний в Internet. - М.:Диалектика, 2005. - 272 с.

5. FileSearchy [Электронный ресурс] // SoftPortal [Офиц. сайт]. URL: http://www.softportal.com/software-33494-filesearchy.html

6. Программа поиска файлов на компьютере FileSearchy [Электронный ресурс] // FileSearchy [Офиц. сайт]. URL: http://www.filesearchy.com/ru/

7. Программа объединения файлов в один документ. [Электронный ресурс] // Блог интернет специалиста [Офиц. сайт]. URL: http://moytop.com/rerajting-i-kopirajting/1-5/bystro-obedinyaem-fajly-word-v-odin-dokument

8. Microsoft Word. [Электронный ресурс] // Википедия [Офиц. сайт]. URL: https://ru.wikipedia.org/wiki/Microsoft_Word

9. Бызова А.К. Развитие системы электронизации Автоматизированного генератора системно обоснованного технического задания путем разработки подсистемы структуризации / А.К. Бызова, С.Л. Гольдштейн, Е.М. Грицюк // Электронный научный журнал "Системная интеграция в здравоохранении". 2016. № 3. С. 5-24. URL: http://sys-int.ru/sites/default/files/sys-int-29-5-24_0.pdf

10. Приказ Минздрава РФ от 17 сентября 1993 г. N 220 «О мерах по развитию и совершенствованию инфекционной службы в Российской Федерации» URL: http://base.consultant.ru/cons/cgi/online.cgi?req=doc;base=LAW;n=100796;dst=0;ts=0E3936A453D9335FBD1B93770135B1AD;rnd=0.980990258282562

11. Душков Б.А. Быстрое чтение // Энциклопедический словарь: Психология труда, управления, инженерная психология и эргономика / Б.А. Душков, А.В. Королев, Б.А. Смирнов – М.: Академический проект; Фонд «Мир», 2005. 848 с.

Грицюк Елена Михайловна, - д.м.н., врач-эпидемиолог ГБУЗ СО ДКБВЛ НПЦ «Бонум», 620149, Екатеринбург, ул. Бардина, 9а, тел: (343)240-42-68 bonum@bonum.info