УДК 004.622

**С.Н. Михайлов,** канд. техн. наук, доцент, ФГБОУ ВО «Юго-Западный государственный университет» (Курск, Россия) (e-mail: rio_kursk@mail.ru)

**О.Е. Ключникова,** канд. техн. наук, доцент, ФГБОУ ВО «Юго-Западный государственный университет» (Курск, Россия) (e-mail: rio_kursk@mail.ru)

# ИНФОЛОГИЧЕСКАЯ СИСТЕМА АНАЛИТИЧЕСКОГО МОНИТОРИНГА ДАННЫХ В НЕСТРУКТУРИРОВАННЫХ ИНФОРМАЦИОННЫХ РЕСУРСАХ

*В работе предлагается вариант решения проблемы оперативного поиска информации в неструктурированных информационных ресурсах. Построены и описаны четыре основных блока, реализующих поиск информации по смысловым значениям. В статье предлагается алгоритм решения задачи оценки соответ-ствия смыслового содержания текстовых документов заданной предметной области. Предложенный инфологический подход выполнен на основе анализа данных патентного поиска, опубликованных научных работ и проведенных экспериментальных исследований эффективных методов автоматической оценки содержания неструктурированных информационных ресурсов для организации процессов информационно-аналитического обеспечения научной деятельности.*

*В работе предложен способ оценки и сопоставления тематической направленности данных в неструктурированных информационных ресурсах, на основе применении инфологической системы. Данный способ предполагает проведение кластеризации текстовых документов путем сравнения семантического содержания исследуемого текста и антологии. Описана структура поисковой подсистемы, имеющей сервисно-ориентированную клиент-серверную архитектуру с тонким клиентом (веб-обозреватель). Описанный метод был апробирован на наборе текстов, полученных в результате мониторинга открытых публичных инфокоммуникационных Интернет-ресурсов без ограничения темы (получено и обработано более 1 млн. экземпляров текстов). Среди полученных текстов экспертным путем была сформирована обучающая выборка для следующих типов текстов: художественные тексты, научные технические статьи, автоматически сгенерированные псевдонаучные тексты, полученные в результате работы систем, спам-содержащие тексты.*

*Предложен состав и описана общая архитектура программного обеспечения инфологической системы, основные компоненты системы являются кросс-платформенными. На основе результатов экспериментальных исследований показана принципиальная возможность реализации автоматизированной оценки тематического подобия документов на примере инфологической обработки текстов рабочих программ дисциплин, сформированы требования, предъявляемые к программному интерфейсу взаимодействия макета с внешними поисковыми системами.*

***Ключевые слова:*** *инфологическая система, оценка тематического подобия, информационный ресурс, рабочая программа дисциплины, компетенция, семантический анализ, смысловое значение.*

**DOI:** 10.21869/2223-1560-2017-21-5-45-61

**Ссылка для цитирования:** Михайлов С.Н., Ключникова О.Е. Инфологическая система аналитического мониторинга данных в неструктурированных информационных ресурсах // Известия Юго-Западного государственного университета. 2017. Т. 21, № 5(74). С. 45-61.

***

В настоящее время все большую актуальность приобретает проблема уменьшения количества не соответствующей запросу информации при информационном поиске, и улучшения правильности запрашиваемого материала. Поиск решения проблем в данной тематике ведется постоянно, но часто «эффект засорения поисковых систем» работает быстрее, чем находятся способы избавиться от этого. Потребность людей в улучшении информационного поиска видна явно, так как пользователям хочется найти интересующую и нужную им информацию как можно быстрее [1, 3]. Однако проведенный анализ литературы [1, 2, 3] показывает, что известные информационно-поисковые системы ориентированы в настоящее время в основном на те функциональные расширения, которые вытекают из возможностей Интернета и компьютерных технологий. Они в

большей степени реализуют компьютерную обработку атрибутивной внешней стороны документов и в недостаточной мере работают с семантическим содержанием текстов. Такие системы недостаточно полно реализуют аналитический мониторинг, а также установление семантической и понятийной эквивалентности текстов.

Причиной этого является отсутствие эффективных методов представления семантико-смыслового содержания текстовых данных в вычислительной среде, что также является серьёзной научной проблемой [4].

Таким образом, проблема оперативного поиска информации в неструктурированных информационных ресурсах и целенаправленного представления тематических данных, релевантных запросу исследователя, в настоящее время является актуальной. Устранение данной проблемы требует решения ряда взаимосвязанных научных задач, ориентированных на разработку и создание систем информационно-аналитического обеспечения научных исследований. Перспективным направлением решения указанных задач является применение в качестве аналитического элемента обеспечения научных исследований инфологической системы аналитического мониторинга.

Одной из основных функций такой системы выступает функция тематической кластеризации документов, обнаруженных в различных информационных ресурсах, с целью оперативного и целенаправленного их доведения до исследователей [1].

Реализация функции кластеризации в инфологической системе требует решения следующих взаимосвязанных задач:

– количественная оценка сходства содержания (тематики) документов;

– количественная оценка принадлежности документа к тематической группе.

Одним из способов проведения количественной оценки сходства и принадлежности является выделение связанных пар ключевых слов в структуре элементарной синтаксической конструкции текста.

Примерами таких пар ключевых слов может служить сочетание слов «связь – мобильный», указывающее на принадлежность текста к области связи, сочетание слов «оперативная память», указывающее на принадлежность текста к области вычислительной техники, и т. д. Построение визуального графа выполняется на основе обработки данных о количестве повторений словосочетаний в тексте, которые хранятся в отдельном XML-файле. Отличительной особенностью инфологической системы является функция хранения антологий требуемых предметных областей (тематик), которые также могут быть представлены визуальным графом.

В рамках научно-исследовательской работы № 13-07-00137 «Исследование и разработка научно-технических путей создания инфологической системы информационно-аналитического обеспечения научных исследований вуза», выполненной при поддержке РФФИ, предложен способ кластеризации текстовых документов на основе сравнения семантического содержания исследуемого текста и антологии. При этом возникает необходимость сравнить результаты анализа исследуемого текста и антологии предметной области, представленные соответствующими XML-файлами.

По результатам данного сравнения должен быть сделан вывод о принадлежности текста к определенной тематической группе. Например, если в эталонном

и исследуемом текстах определенные сочетания пар слов превышают установленный порог, то исследуемый текст можно отнести к предметной области, которую характеризует данная антология (эталон) [1].

Очевидно, для решения данной задачи сравнения текстов необходимо создание механизма обработки документов, обладающего следующими функциональными возможностями [8]:

– наличие структуры данных, в которых хранится информация о ключевых словах, количестве повторений пар ключевых слов в исследуемом и эталонном текстах;

– должен быть реализован механизм поиска одинаковых сочетаний ключевых слов в исследуемом и эталонном текстах. Если данный поиск дал положительные результаты, то далее необходимо выбрать данные, характеризующие количество пар ключевых слов;

– на основе логических правил вида «если – то» должен быть разработан алгоритм, выявляющий, следует ли данный текст отнести к предметной области в соответствии с эталонным текстом.

Поскольку разрабатываемая система должна уметь правильно обрабатывать информацию, полученную от поисковых машин сети Интернет [14] (в нашей работе – это централизованная поисковая система), и возвращать меньшее количество ненужной информации, а в идеальном случае – выдавать только запрашиваемую информацию, то необходимо описать взаимодействие поисковой машины и системы персонализированной подготовки данных с реализацией функции определения тематики и смыслового значения (рис. 1) (в данной работе – субпоисковая система).

Как видно из рисунка 1, персонализированная система подготовки данных фактически является улучшающей надстройкой существующих поисковых машин. Пользователь формирует запрос, который, в свою очередь, обрабатывается определенным образом персонализированной поисковой системой, затем правильно сформированный запрос посылается централизованной поисковой системе.

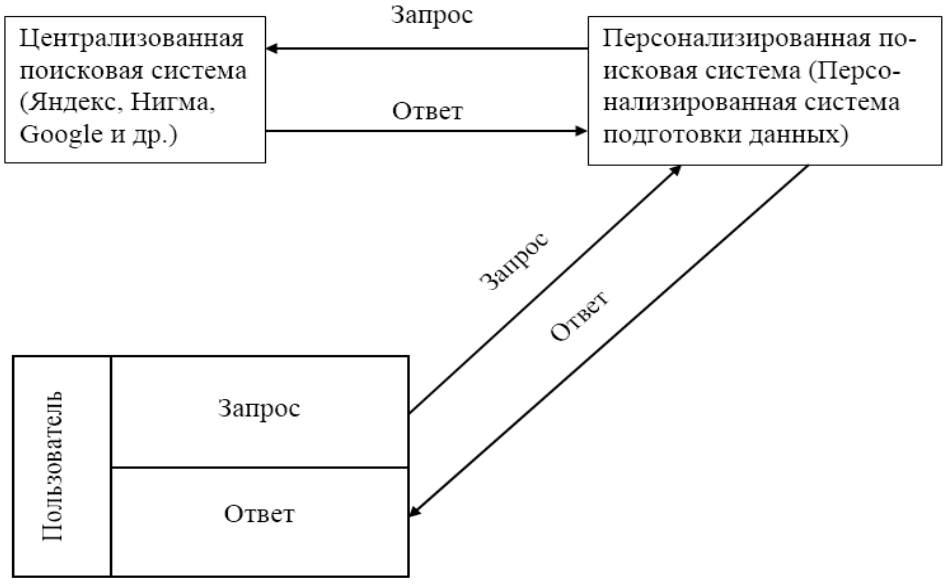


Рис. 1. Обобщенная схема взаимодействия пользователя с персонализированной поисковой системой, которая взаимодействует с централизованной поисковой системой

Основными целями в данной работе являются построение и описание четырех основных блоков, реализующих поиск информации по смысловым значениям. Данные блоки имеют названия: блок технологической подготовки входных информационных ресурсов, блок структуризации и нормализации описания входных информационных ресурсов, блок определения тематики и анализа смысловых значений, блок вывода результатов. Каждый из четырех блоков содержит модули, объединяющие множество программ, алгоритмов и процессов.

Блок технологической подготовки предназначен для качественной оценки семантических отношений между элементами текста и производит первоначальную подготовку файлов различных форматов, выявляет кодировку, классифицирует файлы по форматам [6, 15] и по типам содержащихся в них сообщений (по содержанию), определяет язык [17, 18], на котором сформирован документ, при необходимости имеет функцию переводчика. Конечным результатом функционирования данного блока является исходный файл, переформатированный в единую систему представления (кодирования) [7, 8, 10].

На первом этапе работы блока технологической подготовки входных информационных ресурсов выполняется нормализация текста, представляющая собой приведение его к виду, в котором все слова приведены в нормальную (базовую) форму и исключены стоп-слова (союзы, местоимения и т.д.).

На втором этапе осуществляется сегментация на синтаксические единицы и определение рейтинга связок слов с целью формирования иерархических понятий текста.

На последующих этапах осуществляется процедура наглядного представления иерархии понятий текста путем отображения понятийного графа текста, в узлах которого находятся термины, а дугами обозначаются связи между ними. Таким образом, в результате инфологической обработки документа может быть сформирована понятийная иерархия текста, строго соответствующая семантическому содержанию исходного документа.

Подобная понятийная иерархия представляет собой отражение семантических сущностей текстов, и может быть использована в качестве процесса компактного компьютерного представления семантики текста.

Структурная схема технологической подготовки входных информационных ресурсов приведена на рисунке 2.

Визуальный граф понятийной иерархии, представляемый оператору, является аналогом иероглифической записи документа, позволяющим воспринимать содержимое текста не последовательно, а моментально.

Описанная процедура инфологической обработки текста была положена в основу создания нового способа оценки подобия тематического содержания лекционных курсов путем сравнения понятийных иерархий рабочих программ дисциплин выбранного направления подготовки студентов [5, 6].

В ходе разработки способа инфологической обработки для оценки тематического подобия содержания лекционных курсов был сформирован методологический подход, содержащий 8 этапов [6]:

Этап 1 – Выбор информативных документов, наиболее интенсивно используемых в деятельности кафедры и хранящихся в информационных ресурсах.

Этап 2 – Структурная декомпозиция тематического содержания документов.

Этап 3 – Нумерация документов и содержащихся в них тем.

Этап 4 – Создание архива документов и приведение их к единому формату.

Этап 5 – Инфологическая обработка документов архива на основе формирования упорядоченной совокупности тематических запросов.

Этап 6 – Последовательный анализ признаков обнаружения подобия тематического содержания запроса в имеющихся архивных данных.

Этап 7 – Принятие решения о тематическом сходстве содержания в различных обработанных документах.

Этап 8 – Идентификация тематик, содержащих семантическое подобие в различных документах.

Данная методология легла в основу создания технологии, в которой авторами была определена последовательность операций, необходимых для оценки подобия тематического содержания рабочих программ дисциплин по направлению подготовки 11.03.02 «Инфокоммуникационные технологии и системы связи».

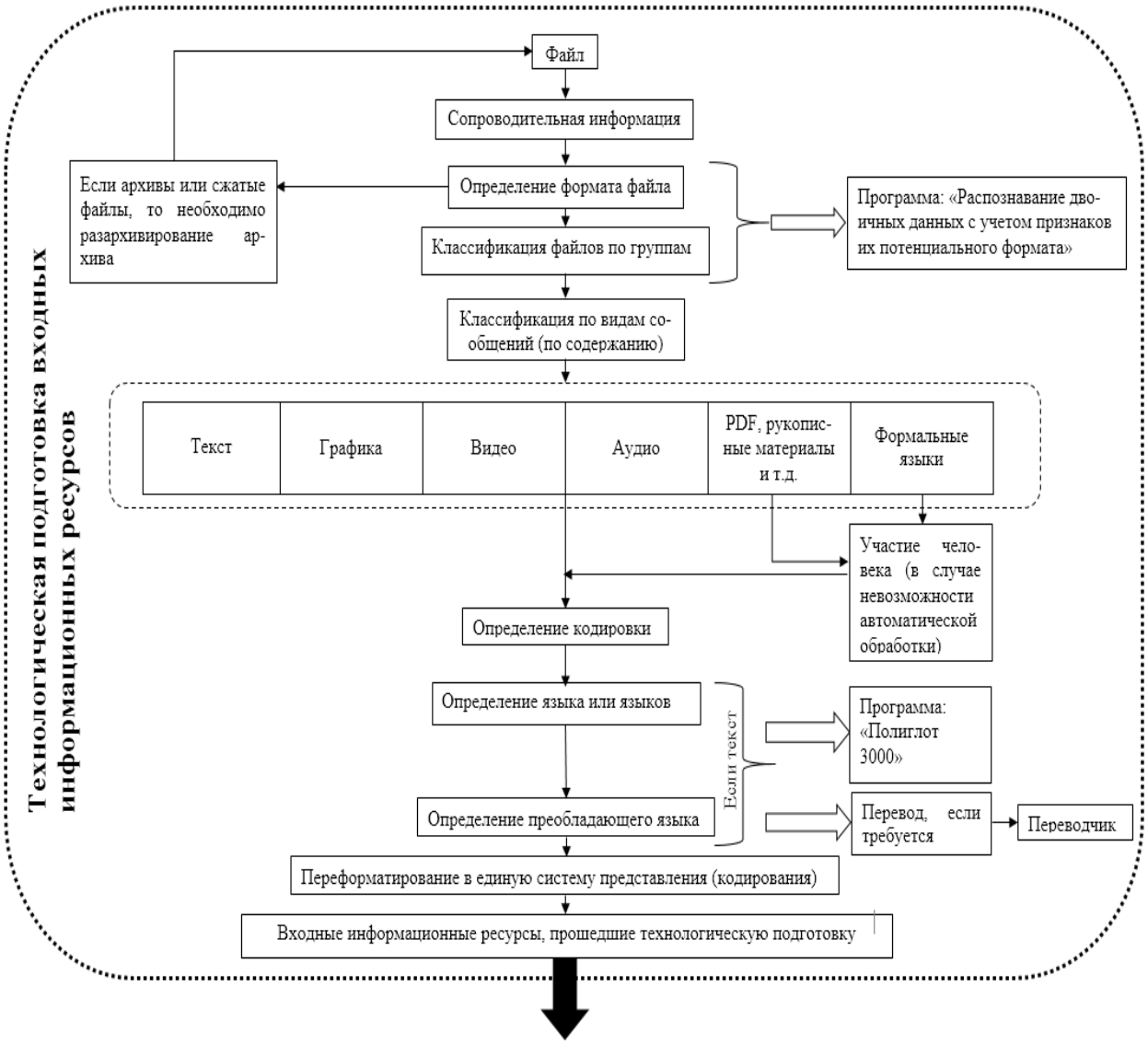


Рис. 2. Структурная схема технологической подготовки входных информационных ресурсов

Апробация разработанной технологии проводилась в ходе выполнения эксперимента, задачей которого была оценка возможности применения инфологической системы для реализации автоматизированного процесса выявления тематического подобия содержания документов, хранящихся в информационных ресурсах. В качестве информационного ресурса была выбрана электронная библиотека сервера выпускающей кафедры [7, 13].

В качестве документов отобраны шестнадцать программ дисциплин профессионального направления подготовки бакалавров кафедры. При этом каждому из N документов был присвоен идентификационный номер (от 1 до 16) для последующего удобства администрирования данных ресурсов ($N_i$, i=1,2…16).

В каждом $N_i$ документе проведена структурная декомпозиция содержащегося в нем тематического материала. В результате каждой теме конкретной дисциплины присвоен оригинальный идентификационный номер$T_{i,j}$, где i – номер дисциплины, j– номер темы, которая изучается в i-й дисциплине. Таким образом был создан нумерованный список тем лекционных материалов по каждой дисциплине. Первая дисциплина содержала 14 тем (номера T1.1 – T1.14), вторая – 6 тем (T2.1 – T2.6), в третьей имелось 5 тем (T3.1 –T3.5) и, наконец, в четвертой – 20 тем (T4.1 – T4.20).

Все 16 полученных документов были сохранены в едином формате с расширением «doc», так как это принципиально при использовании имеющегося макета инфологической системы.

На рисунке 3 представлен сформированный список документов, предназначенных для проведения эксперимента.

| Имя | Дата изменения | Тип | Размер |
|---|---|---|---|
| 1 Беспроводные технологии в РЭС.doc | 07.02.2014 11:29 | Документ Microsoft Word 97-2003 | 29 КБ |
| 2 Методы и средства измерений в телекоммуникационных системах .doc | 07.02.2014 11:30 | Документ Microsoft Word 97-2003 | 29 КБ |
| 3 Оптические цифровые телекоммуникационные системы.doc | 07.02.2014 11:30 | Документ Microsoft Word 97-2003 | 30 КБ |
| 4 Основы теории систем связи с подвижными объектами.doc | 07.02.2014 11:31 | Документ Microsoft Word 97-2003 | 34 КБ |
| 5 Системы и сети связи с подвижными объектами.doc | 07.02.2014 11:31 | Документ Microsoft Word 97-2003 | 31 КБ |
| 6 Современные проблемы науки в области телекоммуникаций.doc | 07.02.2014 11:32 | Документ Microsoft Word 97-2003 | 25 КБ |
| 7 Средства коммутации систем подвижной радиосвязи.doc | 07.02.2014 11:33 | Документ Microsoft Word 97-2003 | 27 КБ |
| 8 Устройства генерирования и формирования сигналов в системах подвижной связи.doc | 07.02.2014 11:33 | Документ Microsoft Word 97-2003 | 37 КБ |
| 9 Устройства приема и обработки радиосигналов в системах подвижной радиосвязи.doc | 07.02.2014 11:34 | Документ Microsoft Word 97-2003 | 34 КБ |
| 10 Электропитание устройств и систем телекоммуникаций.doc | 07.02.2014 11:34 | Документ Microsoft Word 97-2003 | 29 КБ |
| 11 Теория электрической связи.doc | 07.02.2014 11:35 | Документ Microsoft Word 97-2003 | 57 КБ |
| 12 Синхронные и плезиохронные цифровые телекоммуникационные системы связи.doc | 07.02.2014 11:35 | Документ Microsoft Word 97-2003 | 25 КБ |
| 13 Теория телекоммуникационных систем и сетей.doc | 07.02.2014 11:35 | Документ Microsoft Word 97-2003 | 33 КБ |
| 14 Менеджмент в телекоммуникациях.doc | 07.02.2014 11:36 | Документ Microsoft Word 97-2003 | 28 КБ |
| 15 Метрология, стандартизация и сертификация.doc | 07.02.2014 11:36 | Документ Microsoft Word 97-2003 | 31 КБ |
| 16 Электромагнитные поля и волны.doc | 07.02.2014 11:37 | Документ Microsoft Word 97-2003 | 27 КБ |

Рис. 3. Нумерованный список документов, участвующих в эксперименте

На очередном этапе эксперимента сформирован архив документов, который будет использоваться непосредственно инфологической системой для смысловой обработки.

Загрузив архив из 16 документов в программный макет инфологической системы, пользователь получает возможность отображения семантических сущностей текста в виде понятийного окру-

жения каждого документа посредством визуального графа понятийной иерархии.

Предложенный подход позволяет сравнивать различные тексты между собой на основе их семантической составляющей, путем визуализации семантического окружения текстов – отображения документа в виде понятийного графа, отображающего связи наиболее значимых понятий (связок элементов), используемых в документе. Данный способ позволяет минимизировать затраты времени на ознакомление с содержанием текста.

Для построения графов изначально строятся рейтинговые распределения связей элементов. Считается, что между каждой парой таких элементов, встречающихся в одном предложении, имеется связь с рейтингом, равным единице. Для каждой связи запускается счетчик, учитывающий, сколько раз та или иная связка элементов предложения встречается в тексте. Если связки элементов встречаются повторно, то значение счетчика увеличивается на единицу. Таким образом, обрабатываются все связи всех предложений текста и подсчитываются их рейтинги.

На рисунке 4 показан вариант визуального графа понятийной иерархии.


Рис. 4. Визуализация семантического окружения текста для документа «Беспроводные технологии в РЭС»

После подсчета рейтингов связей элементов из результата исключаются связи с рейтингом меньше двух и отдельные слова, оставшиеся вообще без связей. В результате формируется список слов, двоек слов и троек слов, и производится ранжирование, т.е. упорядочение списков элементов по убыванию рейтинга соответствующих им связей. Далее на основе списка слов, двоек слов и троек слов текста, полученных на предыдущем этапе метода, происходит формирование понятий, т.е. выделение из списка слов и связок слов тексте понятий, соответствующих заданной предметной области с использованием тезауруса и словарей предметной области и сохранение результирующего массива в формате XML.

Благодаря гибкости и строгости формата XML он часто используется для хранения и передачи данных.

На последующих операциях этапа сформированные запросы последовательно вводились в окно поиска инфологической системы, после чего производилась смысловая обработка всех документов архива на предмет оценки тематического подобия запроса в содержании каждого из документов. Документы, в которых обнаружено семантическое сходство с тематикой запроса, представлялись в виде перечня в окне поиска, как это показано на рисунке 5.

Выполняя последовательный поиск по каждому запросу, пользователь получает возможность определить полный перечень дисциплин, в которых освещается та или иная тема по всему направлению подготовки.

Анализ представленных результатов показывает, что отдельные темы имеют смысловое подобие в 7 и более дисциплинах. Указанный результат подтверждает тот факт, что отдельные темы одной дисциплины в ходе учебного процесса могут иметь многократное повторение в нескольких других дисциплинах.

На рисунках 6, 7 графически представлены обобщенные результаты эксперимента, отражающие количественные и качественные характеристики обнаруженных повторений.

Если в качестве критерия возможных повторений выбрать пять и более совпадений тематического подобия, то получим перечень из 21 рассмотренных тем, представленных на рисунке 6. По этим темам, используя полученные данные, достаточно просто определяется перечень дисциплин, содержащих тематическое подобие, как это показано на рисунке 7.

Выявленные по тематическому подобию дисциплины могут быть представлены руководству кафедры для последующего анализа их тематического содержания и уточнения изучаемых вопросов с целью исключения возможности значительных повторений дидактических единиц. При этом указанная информация может быть использована для поддержки принятия организационных и управленческих решений, направленных на оптимизацию использования учебного времени.

Оценка временных затрат, необходимых для проведения анализа возможных тематических дублирований в 16 РПД ручным способом и автоматизировано с применением инфологической обработки, показывает, что во втором случае оперативность достижения результата может быть повышена до пяти раз [6].

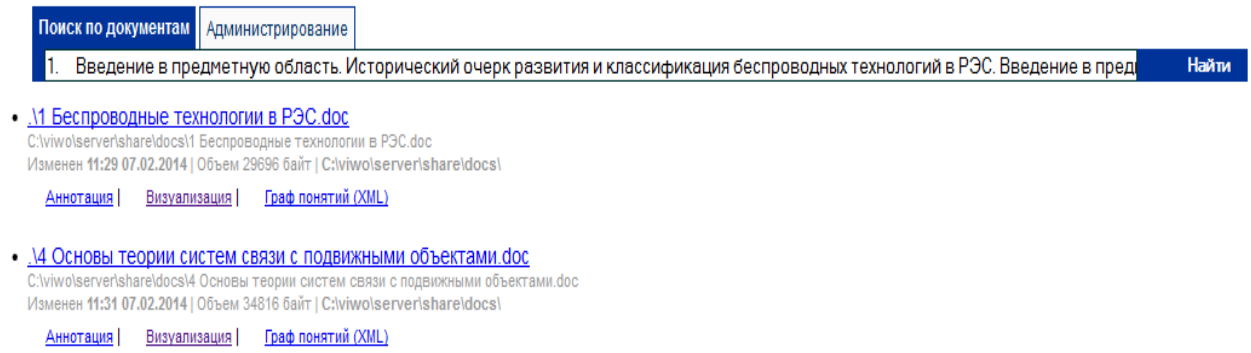


Рис. 5. Результат тематического запроса на предмет семантического подобия

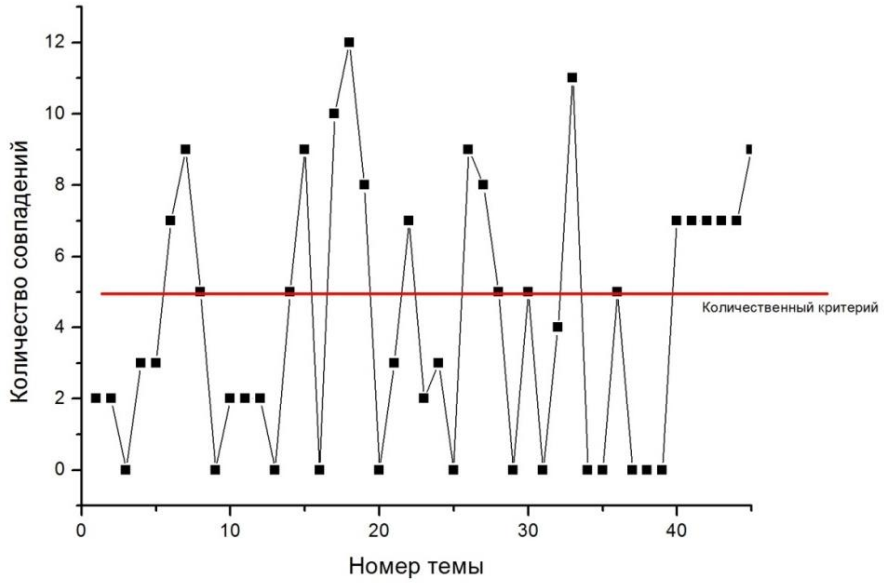


Рис. 6. Количество обнаруженных смысловых подобий тем в разных дисциплинах

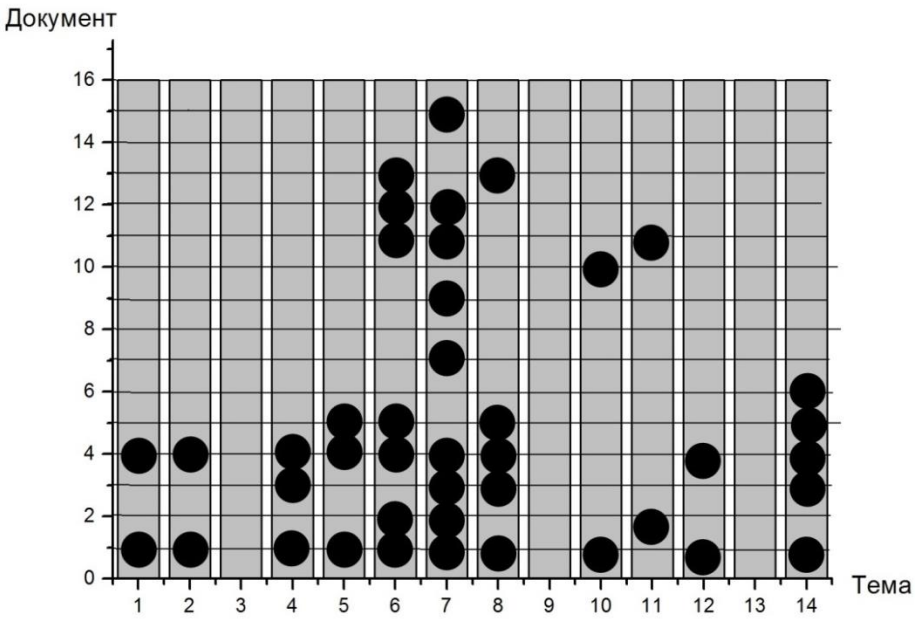


Рис. 7. Отображение возможных повторений отдельных тем в содержании изучаемых дисциплин

В блоке структуризации и нормализации описания с помощью различных методов реализовано упорядоченное описание исходного переформатированного в единую систему представления файла. На выходе данного блока мы получаем файл в установленном структурированном виде.

Практическая реализация метода нормализации текстового документа, а именно - частичная реализация блока структуризации и нормализации описания выполнена в виде программного продукта, названного «Нормализатор первичных данных», реализованного на языке программирования Delphi. Интерфейс программного модуля представлен на рисунке 8.

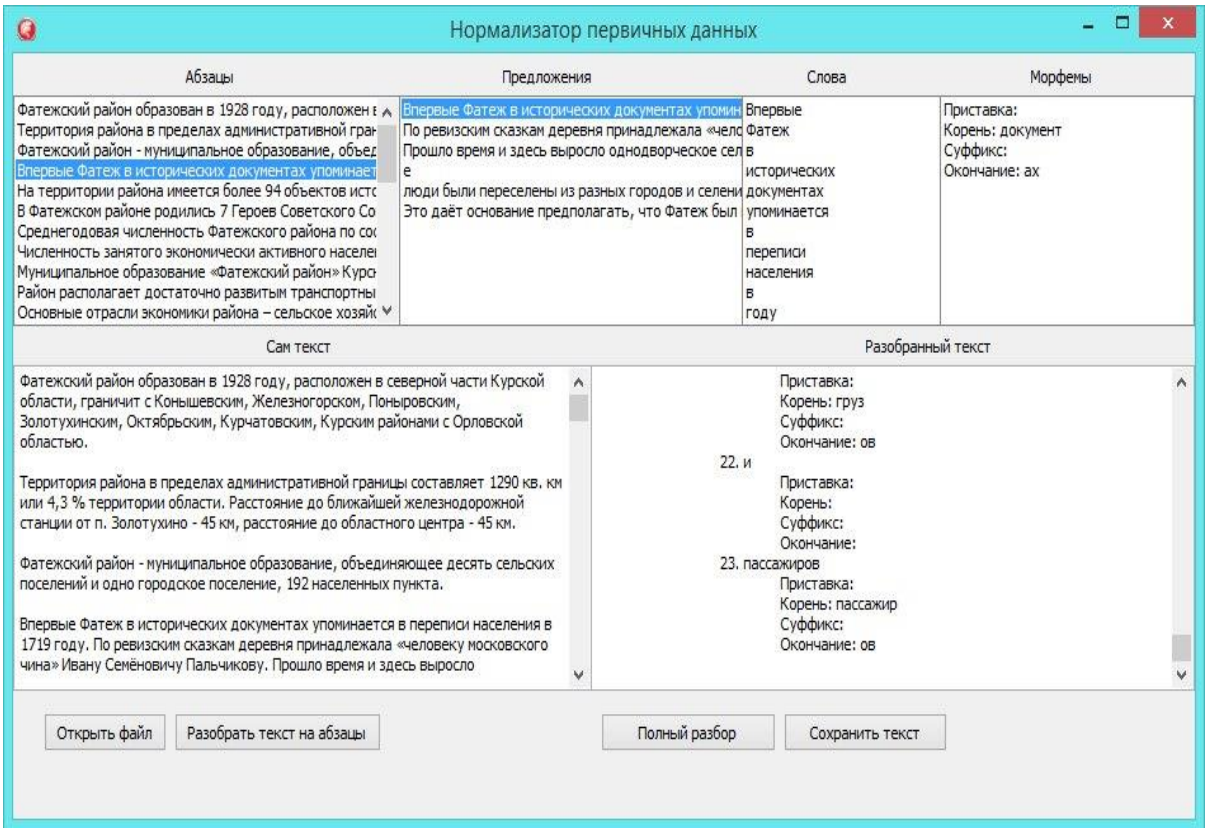


Рис. 8. Интерфейс нормализатора первичных данных

Перед началом инфологической обработки документа проводится его разбор. Разбор текста на его составляющие части (абзацы, предложения, слова, слова разложить в морфемы), выполняется путем помещения в специально отведенный каталог (папку) [6, 11]. Далее при выполнении функции «Разобрать текст на абзацы» происходит определение специальных символов абзаца и разложение текста на абзацы. При выполнении функции «Полный разбор» происходит структурное разложение всех имеющихся абзацев на более мелкие структурные элементы текста, а именно: абзацы раскладываются в предложения, предложения в слова, слова, в свою очередь, проходят морфемный разбор.

На рисунке 9 показан вид окна управления и отображения программного модуля нормализатора первичных данных.

Исходный текст загружается в область с номером 8. Для загрузки текста в область 8 необходимо нажать кнопку 10, при этом произойдет внесение исходного текста в область 8 и разбор текста на абзацы с внесением всех абзацев в область 1. Далее нажатием кнопки 7 происходит полное структурное разложение исходного текста в область 5. У данного программного продукта имеется функция работы с отдельными элементами. Для этого необходимо выбрать нужный абзац нажатием левой кнопки мыши в области 1. После чего выбранный абзац переносится в область 2, где происходит разложение абзаца на предложения. В области 2 такими же операциями, как и в области 1, выбираем нужное предложение и нажимаем левую кнопку мыши, после чего в области 3 отображаются все слова выбранного предложения.

Далее при необходимости просмотра морфемного разбора какого-то слова из области 3 следует выбрать нужное слово и нажать левой кнопкой мыши на данное слово, после чего в области 4 происходит морфемный разбор выбранного слова. Также есть немаловажная функция сохранить результат нормализации из области 5 в файл нажатием кнопки 6. Сохранение в файл позволяет нам в дальнейшем нормализованный документ передать следующим обрабатывающим блокам [11].

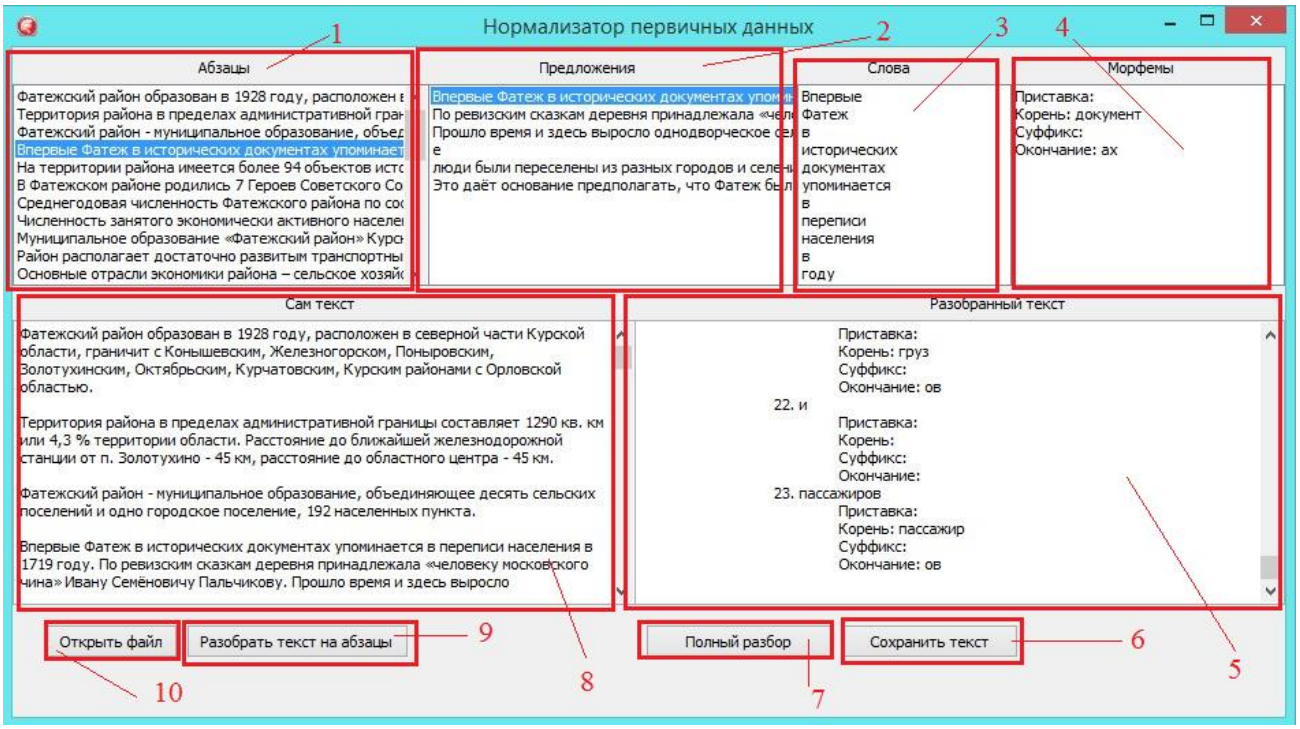


Рис. 9. Области управления и отображения нормализатора первичных данных

Разработанный программный модуль является адаптируемой под необходимые задачи.

Третий блок системы реализует определение тематики, анализ смысловых значений, сравнение текстов на основе их смысловых значений. Данный блок содержит модуль принятия решений, который решает, сходятся тексты по смыслу или нет.

Инструмент определения тематики и анализа смысловых значений состоит из четырех модулей:

– определение тематики;

– анализатор смыслового значения;

– блок сравнения смысловых значений (СЗ);

– блок принятия решений.

После того, как исходный файл прошел технологическую подготовку [5, 12], структуризацию и нормализацию, обработанный и стандартизованно-структурированный файл подвергается дальнейшей обработке для определения тематики и СЗ. Определив СЗ одного текста, система производит сравнение различных текстов на основе СЗ. После выполнения операции сравнения пользователю через блок вывода информации выдается результат поиска. В анализаторе смысло-

вых значений реализованы два метода определения СЗ:

– определение СЗ на основе различных словарей;

– описание слов и соответственно их СЗ на основе различных наук (физики, химии, биологии, технических наук и многих других).

Для реализации метода на основе словарей в систему необходимо загрузить уже существующие словари или внести собственноручно необходимые термины и определения. В настоящее время существует множество электронных словарей различной направленности (медицинские словари, технические словари, толковые словари и т.п.).

На рисунке 10 представлена схема определения СЗ.

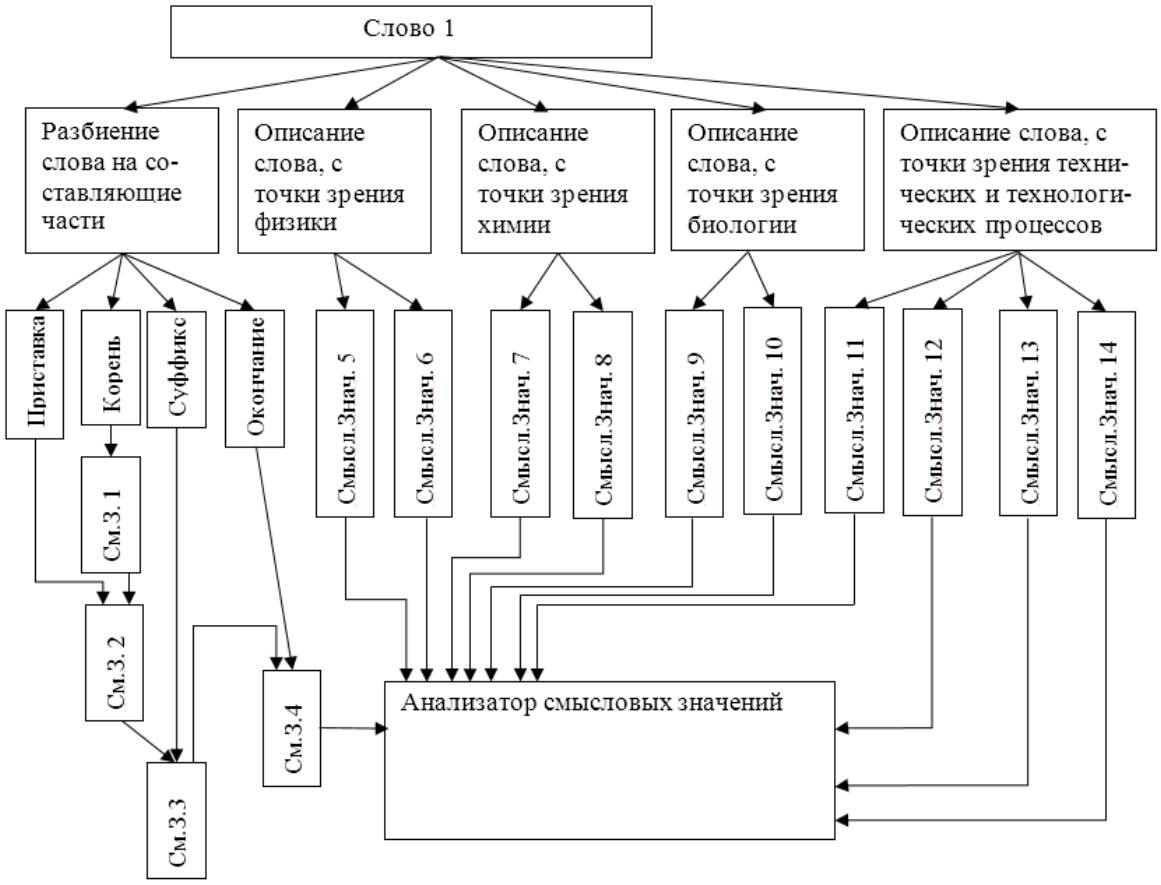


Рис. 10. Обобщенная структурная схема анализатора смысловых значений

По завершении процесса определения СЗ начинается операция сравнения текстов по смысловым значениям и процедура принятия решения о сходстве текстов. В данной реализации сравнение текстов осуществляется в блоке сравнения смысловых значений, а затем в блоке принятия решений формируется вариант решения и результата. Первоначальная задача – сравнить смысл у слов, затем блок принятия решений определяет, совпадают смыслы у слов или нет. В случае совпадения блок принятия решений отдает управляющую команду блоку сравнения о расширении сравнения СЗ, а именно сравнение смыслов расширяется от слов до словосочетаний. Сравнив смыслы словосочетаний, блок сравнения вновь передает

обработанную информацию блоку принятия решений, при совпадении СЗ сравнение расширяется до предложений, абзацев и полного текста. В случае если СЗ не совпадают, то процесс сравнения останавливается и блок принятия решений производит удаление информации, не соответствующей смыслу запроса.

На рисунке 11 представлена схема работы блока сравнения и блока принятия решений.

В конечном итоге система аналитического мониторинга, проделав полный цикл от технологической подготовки данных до оценки и сравнения СЗ, выдаст результат сходства текстов, удалив из каталога файлы, не совпадающие по СЗ, возможно предложит расширить поиск. Также система должна сообщить пользователю о случаях, когда она не может без участия человека обработать документ.

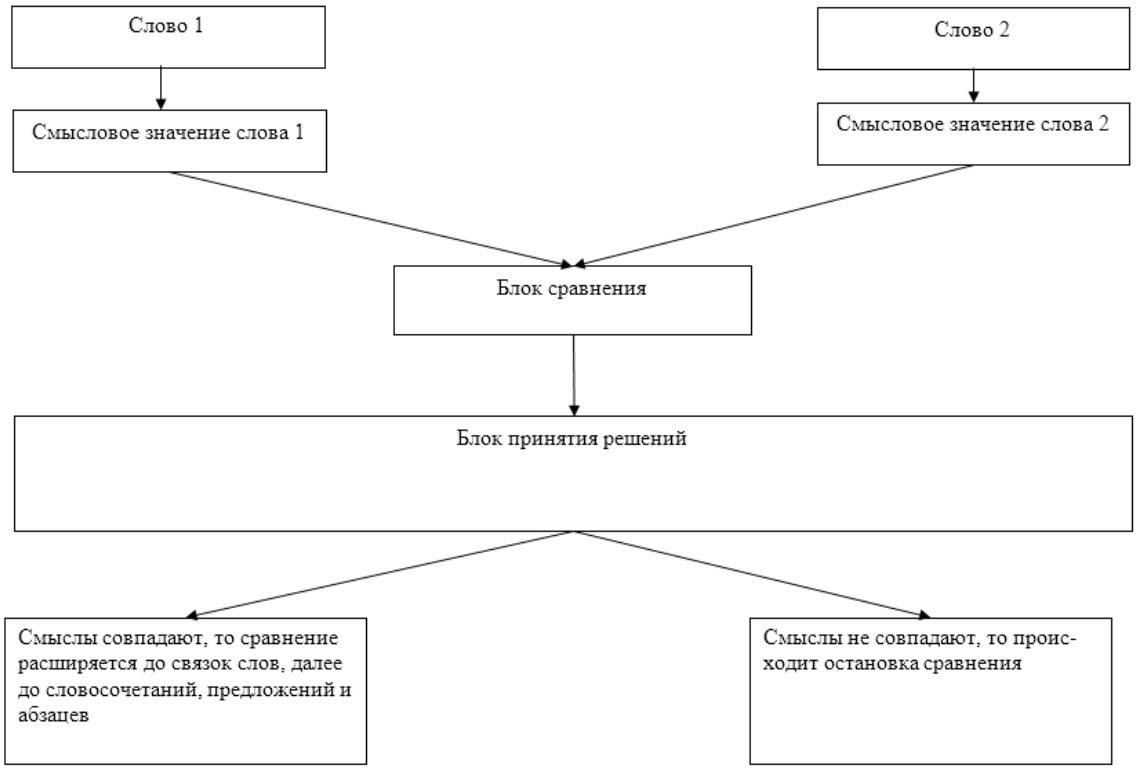


Рис. 11. Процедура сравнения двух слов на основе СЗ: слово 1- слово запроса; слово 2 – искомое слово (возможно, заданное другими буквами, но имеющее схожий смысл)

Четвертый блок получает информацию от блока принятия решений и выводит ее пользователю через соответствующий интерфейс [15].

**Заключение.** Описанная в работе инфологическая система аналитического мониторинга данных неструктурированных информационных ресурсов, включающая четыре основных блока, позволяет реализовать тематический поиск информации на основе оценки семантико-смысловой близости текстовых документов.

В работе показано информационное взаимодействие блоков технологической подготовки входных информационных ресурсов, структуризации и нормализации описания, определения тематики и анализа смысловых значений и вывода

результатов в едином технологическом цикле инфологической обработки текстовых документов.

Приведенные данные экспериментальных исследований предлагаемой системы аналитического мониторинга данных в информационных ресурсах для оценки повторяемости изучаемых тем в дисциплинах отдельного направления подготовки студентов позволили оценить повышение оперативности решения задачи до пяти раз по сравнению с традиционным (ручным) способом ее решения.

Наиболее перспективным направлением применения подобных информационных систем обработки данных в неструктурированных информационных ресурсах является семантический поиск требуемых тематических данных, упреждающее информационное обеспечение исследователей и научных коллективов новыми данными по установленным тематикам в процессе организации научно-исследовательской и педагогической деятельности.

**Список литературы**

1. Михайлов С.Н. Способ тематической кластеризации текстовых документов на основе их инфологической обработки // Наукоемкие технологии. 2012. Т. 13, № 9. С. 48-51.

2. Кулешов С.В., Михайлов С.Н. Вариант архитектуры субпоисковой системы для реализации функции аналитического мониторинга // Труды СПИИРАН. 2013. № 8 (31). С. 247-254

3. Зайцева А.А., Кулешов С.В., Михайлов С.Н. Метод оценки качества текстов в задачах аналитического мониторинга информационных ресурсов// Труды СПИИРАН. 2014. № 6. С. 144-155.

4. Михайлов С.Н., Кулешов С.В. Экспертный мониторинг неструктурированных информационных ресурсов в интересах информационно-аналитического обеспечения космических исследований// Известия Юго-Западного государственного университета. 2013. № 6-2 (51). С. 40-43.

5. Михайлов С.Н., Агапченко К.И. Способ инфологической обработки рабочих программ дисциплин для оценки подобия тематического содержания лекционных курсов // Инфокоммуникации и информационная безопасность: состояние, проблемы и пути решения: материалы I Всероссийской научно-практической конференции. Курск, 2014. С. 128-136.

6. Михайлов С.Н., Чуйкова В.В. Способ оценки содержания дисциплин отдельного направления подготовки требуемым компетенциям // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2014. № 3. С. 19-24.

7. Михайлов С.Н., Хотынюк С.С., Потапенко А.М. Технологии интерактивного выявления смыслового содержания текстов в целях организации информационно-аналитического обеспечения научных исследований // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2013. № 4. С. 29-34.

8. Михайлов С.Н., Тезик К.А. Вариант программной реализации способа тематической кластеризации текстовых документов на основе использования макросов VBA и EXCEL // Известия Юго-Западного государственного университета. 2012. № 4 (43), ч.2. С. 17-21.

9. Михайлов С.Н., Севрюков А.Е. Обобщенная архитектура инфокоммуникационной среды информационно-аналитического обеспечения научных исследований вуза // Информационно-измерительные и управляющие системы. 2010. Т. 8, № 11. С. 40-42.

10. Марухленко А.Л., Конарев Д.И., Якушев А.С. Сравнение текстов на основе анализа и сопоставления их смысловых значений // Инфокоммуникации и информационная безопасность: состояние, проблемы и пути решения: материалы II Всероссийской научно-практической конференции. Курск, 2015. С.168-171.

11. Марухленко А.Л., Коршунов Е.Е., Якушев А.С. Вариант нормализации первичных данных с учетом семантической составляющей // Инфокоммуникации и информационная безопасность: состояние, проблемы и пути решения: материалы II Всероссийской научно-практической конференции. Курск, 2015. С.171-176.

12. Потапенко А.М., Юрченко А.Г., Попадинец Р.В. Семиотическая модель языка // Нейрокомпьютеры: разработка, применение. 2014. № 6. С. 34-41

13. Исследование и разработка научно-технических путей создания информационно-телекоммуникационной системы аналитического мониторинга в неструктурированных информационных ресурсах: отчет о НИР / Юго-Зап. гос. ун-т (ЮЗГУ); рук. М.В. Соколова. Курск, 2015. 293 с. № 2.2491.2014/К.

14. Тезик К.А., Михайлов С.Н. Методика планирования эксперимента в целях распознавания тематической направленности информационных ресурсов сети интернет // Инфокоммуникации и информационная безопасность: состояние, проблемы и пути решения: материалы II Всероссийской научно-практической конференции. Курск, 2015. С.72-79.

15. Классификация форматов файлов для задач селекции документов / А.С. Якушев [и др.] // Инфокоммуникации и информационная безопасность: состояние, проблемы и пути решения: материалы I Всероссийской научно-практической конференции. Курск, 2014. С. 289-293.

16. Потапенко А.М., Русанов Р.Н. Проблема информационного поиска по содержанию // Известия Юго-Западного государственного университета. Серия Управление, вычислительная техника, информатика. Медицинское приборостроение. 2012. № 2, ч.3. С. 100-102.

17. Потапенко А.М., Юрченко А.Г. Схема образования языковых знаков в естественно-языковых текстах // Нейрокомпьютеры: разработка, применение. 2014. № 6. С. 41-44.

18. Персонализированная система поиска информации с функцией определения тематики и анализа смысловых значений / А.М. Потапенко, А.Л. Марухленко, Д.И. Конарев, А.С. Якушев // Инфокоммуникации и информационная безопасность: состояние, проблемы и пути решения: материалы II Всероссийской научно-практической конференции. Курск, 2015. С. 181-187.

*Поступила в редакцию 14.08.17*

---

UDC 004.622

**S.N. Mikhailov,** Candidate of Engineering Sciences, Associate Professor**,** Southwest State University (Kursk, Russia) (e-mail: rio_kursk@mail.ru)

**O.E. Klyuchnikova,** Candidate of Engineering Sciences, Associate Professor**,** Southwest State University (Kursk, Russia) (e-mail: rio_kursk@mail.ru)

**INFOLOGICAL MONITORING SYSTEM OF ANALYTICAL DATA UNSTRUCTURED CONTENT**

*In operation the way of solving the problem of quick search of information in unstructured information resources is offered. Four main units realizing information search in semantic values are constructed and described. In article the algorithm of the decision of the task of assessment of compliance of semantic contents of text documents of the*

*given data domain is offered. The offered infologichesky approach is executed on the basis of data analysis of patent search, the published scientific operations and the conducted pilot studies of effective methods of automatic assessment of maintenance of unstructured information resources for the organization of processes of information and analytical support of scientific activities.*

*In operation the method of assessment and comparison of a subject directivity of data in unstructured information resources, on a basis use of infologichesky system is offered. This method assumes carrying out a clustering of text documents by comparing of semantic contents of the researched text and the anthology. The structure of the retrieval subsystem having the service-oriented client-server architecture with the thin client (web observer) is described. The described method was approved on a set of the texts received as a result of monitoring of open public infocommunication Internet resources without restriction of a subject (more than 1 million copies of texts are received and processed). Among the received texts by an expert way learning selection for the following types of texts was created: artistic texts, scientific technical articles, the pseudoscientific texts received as a result of operation of systems, a spam automatically generated - the containing texts.*

*The composition is offered and the general architecture of the software of infologichesky system is described, principal components of system are cross-platform. On the basis of results of the pilot studies the basic possibility of implementation of automated assessment of subject similarity of documents on the example of infologichesky processing of texts of working programs of disciplines is shown, requirements imposed to the program interface of interaction of a prototype with external search engines are created.****Key words:*** *infological system, assessment of the thematic similarity, information resource working program of discipline, competence, semantic analysis, meaning.*

**DOI:** 10.21869/2223-1560-2017-21-5-45-61

**For citation:** Mikhailov S.N., Klyuchnikova O.E., Infological Monitoring System Of Analytical Data Unstructured Content. Proceedings of Southwest State University, 2017, vol. 21, no. 5(74), pp. 45-61 (in Russ.).

***

**Reference**

1. Mihajlov S.N. Sposob tematicheskoj klasterizacii tekstovyh dokumentov na osnove ih infologicheskoj obrabotki // Naukoemkie tehnologii, 2012, vol. 13, no. 9, pp. 48-51.

2. Kuleshov S.V., Mihajlov S.N. Variant arhitektury subpoiskovoj sistemy dlja realizacii funkcii analiticheskogo monitoringa. Trudy SPIIRAN. 2013, no. 8 (31), pp. 247-254.

3. Zajceva A.A., Kuleshov S.V., Mihajlov S.N. Metod ocenki kachestva tekstov v zadachah analiticheskogo monitoringa informacionnyh resursov. Trudy SPIIRAN, 2014, no. 6, pp. 144-155.

4. Mihajlov S.N., Kuleshov S.V. Jekspertnyj monitoring nestrukturiro-vannyh informacionnyh resursov v interesah informacionno-analiticheskogo obespechenija kosmicheskih issledovanij. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta, 2013, no. 6 (51), pt. 2, pp. 40-43.

5. Mihajlov S.N., Agapchenko K.I. Sposob infologicheskoj obrabotki rabochih programm disciplin dlja ocenki podobija tematicheskogo soderzhanija lekcionnyh kursov. Infokommunikacii i informacionnaja bezopasnost': sostojanie, problemy i puti reshenija: materialy I Vserossijskoj nauchno-prakticheskoj konferencii. Kursk, 2014, pp. 128-136.

6. Mihajlov S.N., Chujkova V.V. Sposob ocenki soderzhanija disciplin otdel'nogo napravlenija podgotovki trebuemym kompetencijam. Izvestija Jugo-Zapadnogo gosudarstvennogo universite-ta. Serija: Upravlenie, vychislitel'naja tehnika, informatika. Medicinskoe priborostroenie, 2014, no. 3, pp. 19-24.

7. Mihajlov S.N., Hotynjuk S.S., Potapenko A.M. Tehnologii interaktiv-nogo vyjavlenija smyslovogo soderzhanija tekstov v celjah organizacii informaci-onno-analiticheskogo obespechenija nauch-nyh issledovanij. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta. Serija: Uprav-

lenie, vychislitel'naja tehnika, informatika. Medicinskoe priborostroenie, 2013. no. 4, pp. 29-34.

8. Mihajlov S.N., Tezik K.A. Variant programmnoj realizacii sposoba tematicheskoj klasterizacii tekstovyh dokumentov na osnove ispol'zovanija makrosov VBA i EXCEL. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta, 2012, no. 4 (43), pt. 2, pp. 17-21.

9. Mihajlov S.N., Sevrjukov A.E. Obobshhennaja arhitektura infokommunikacionnoj sredy informacionno-analiticheskogo obespechenija nauchnyh issledovanij vuza. Informacionno-izmeritel'nye i upravljajushhie sistemy, 2010, vol. 8, no. 11, pp. 40-42.

10. Maruhlenko A.L., Konarev D.I., Jakushev A.S. Sravnenie tekstov na osnove analiza i sopostavlenija ih smyslovyh znachenij. Infokommunikacii i informacionnaja bezopasnost': sostojanie, problemy i puti reshenija: materialy II Vserossijskoj nauchno-prakticheskoj konferencii. Kursk, 2015, pp.168-171.

11. Maruhlenko A.L., Korshunov E.E., Jakushev A.S. Variant normalizacii pervichnyh dannyh s uchetom semanticheskoj sostavljajushhej. Infokommunikacii i informacionnaja bezopasnost': sostojanie, problemy i puti reshenija: materialy II Vserossijskoj nauchno-prakticheskoj konferencii. Kursk, 2015, pp.171-176.

12. Potapenko A.M., Jurchenko A.G., Popadinec R.V. Semioticheskaja model' jazyka. Nejrokomp'jutery: razrabotka, primenenie, 2014, no. 6, pp. 34-41

13. Issledovanie i razrabotka nauchno-tehnicheskih putej sozdanija informacionno-telekommunikacionnoj sistemy analiticheskogo monitoringa v nestrukturirovannyh informacionnyh resursah: otchet o NIR; Jugo-Zap. gos. un-t (JuZGU); ruk. M.V. Sokolova. Kursk, 2015. 293 p. No 2.2491.2014/K.

14. Tezik K.A., Mihajlov S.N. Metodika planirovanija jeksperimenta v celjah raspoznavanija tematicheskoj napravlennosti informacionnyh resursov seti internet. Infokommunikacii i informacionnaja bezopasnost': sostojanie, problemy i puti reshenija: materialy II Vserossijskoj nauchno-prakticheskoj konferencii. Kursk, 2015, pp.72-79.

15. Jakushev A.S. [i dr.]Klassifikacija formatov fajlov dlja zadach selekcii dokumentov. Infokommunikacii i informacionnaja bezopasnost': sostojanie, problemy i puti reshenija: materialy I Vserossijskoj nauchno-prakticheskoj konferencii. Kursk, 2014, pp. 289-293.

16. Potapenko A.M., Rusanov R.N. Problema informacionnogo poiska po soderzhaniju. Izvestija Jugo-Zapadnogo gosudarstvennogo universiteta. Serija: Upravlenie, vychislitel'naja tehnika, informatika. Medicinskoe priborostroenie, 2012, no. 2, pt. 3, pp. 100-102.

17. Potapenko A.M., Jurchenko A.G. Shema obrazovanija jazykovyh znakov v estestvenno-jazykovyh tekstah. Nejrokomp'jutery: razrabotka, primenenie, 2014, no. 6, pp. 41-44.

18. Potapenko A.M., Maruhlenko A.L., Konarev D.I., Jakushev A.S. Personalizirovannaja sistema poiska informacii s funkciej opredelenija tematiki i analiza smyslovyh znachenij. Infokommunikacii i informacionnaja bezopasnost': sostojanie, problemy i puti reshenija: materialy II Vserossijskoj nauchno-prakticheskoj konferencii. Kursk, 2015, pp. 181-187.