

На правах рукописи



Толкунов Александр Александрович

**МОДЕЛЬ ОПЕРАТИВНОЙ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ
ТЕКСТОВЫХ КОММЕНТАРИЕВ К ЗАКОНОПРОЕКТАМ**

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Орёл – 2014

Работа выполнена в Академии ФСО России.

Научный руководитель:

Кукушкин Александр Антонович,
кандидат технических наук, доцент

Официальные оппоненты:

Хорошилов Александр Алексеевич,
доктор технических наук,
Федеральное государственное бюджетное учреждение науки Институт проблем информатики Российской академии наук, г. Москва,
ведущий научный сотрудник

Золотарёв Олег Васильевич,
кандидат технических наук, доцент,
Негосударственное образовательное учреждение высшего профессионального образования "Российский новый университет", г. Москва,
доцент

Ведущая организация:

Федеральное государственное бюджетное учреждение науки Институт системного анализа Российской академии наук, г. Москва

Защита состоится "4" июня 2014 года в 16 часов 00 минут на заседании диссертационного совета Д 002.073.01 на базе Федерального государственного бюджетного учреждения науки Институт проблем информатики Российской академии наук (ИПИ РАН) по адресу: 119333, г. Москва, ул. Вавилова, д. 44, корп. 2.

С диссертацией можно ознакомиться в библиотеке и на сайте ИПИ РАН, www.ipiran.ru.

Автореферат разослан "25" апреля 2014 г.

Ученый секретарь
диссертационного совета Д 002.073.01
доктор технических наук, профессор



С.Н. Гринченко

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. В настоящее время в соответствии с Постановлением Правительства РФ от 25.08.2012 № 851 разрабатываемые законопроекты должны проходить процедуру общественного обсуждения на интернет-портале regulation.gov.ru. Результатом такой процедуры является массив комментариев, содержащих предложения интернет-пользователей по корректировке законопроекта. Комментарии подлежат экспертной обработке с целью изучения и обобщения поступивших предложений. Трудоемкость данной процедуры на практике оказывается значительной, поскольку число комментариев для законопроектов, вызвавших наибольшее обсуждение, может достигать 10–20 тысяч. В этих условиях возникает потребность в разработке моделей обработки комментариев, направленных на эффективное ознакомление с поступившими предложениями.

Одним из основных понятий при исследовании процедур анализа комментариев является понятие "мнение". Исследованию способов обработки мнений посвящено научное направление Opinion Mining (*англ.*: анализ мнений), входящее в состав Text Mining (*англ.*: глубинный анализ текстов). Анализ работ Б. Пэнга, Л. Ли, М. Геймона, Дж. Каренини, М. Девалли, П. Резника, М. Ху, Т. Пинча, Т. Муллена и др., посвященных данному направлению исследований, показал, что в основном авторы понимают анализ мнений как задачу классификации по тональности и поиска ключевых слов, в то же время вопросы, касающиеся эффективного ознакомления с мнениями, направленными на улучшение объекта обсуждения, остаются недостаточно исследованными.

Одним из подходов к моделированию процедур обработки мнений может стать подход, учитывающий многомерность представления мнения. Для обработки многомерных данных одним из наиболее эффективных методов обработки является оперативная аналитическая обработка данных, предложенная Э. Ф. Коддом. Модификации данного метода обработки для текстов описаны в трудах Дж. Мозе, С. Кейта, Х. Ло, С. Лин, Дж. Хана, Д. Жана, Й. Ю, Б. Джанета, А. Инокучи, К. Такеда и др. Однако анализ данных работ показывает, что предлагаемые модели не универсальны, авторы вводят иерархии категорий, актуальные в конкретной задаче, а вновь вводимые операции погружения ограничивают лишь списками ключевых слов.

В данной ситуации возникли условия для постановки и решения **научной задачи** по разработке модели эффективной обработки комментариев с использованием оперативной аналитической обработки текстов. Разработанная модель обработки комментариев обеспечивает поддержку рекурсивной процедуры понимания комментариев экспертом за счет последовательного перехода между классами комментариев с погружением и ознакомлением с предложениями по корректировке законопроекта по дедуктивной схеме: интенсивность мнений в классе комментариев – ознакомление с типичными комментариями – ознакомление с полезными комментариями – ознакомление со всеми комментариями.

Объект исследования: процесс аналитической обработки комментариев интернет-пользователей в ходе общественного обсуждения законопроектов.

Предмет исследования: модели оперативной аналитической обработки текстов и алгоритмы погружения в текстовые данные в ходе обработки результатов общественного обсуждения законопроектов.

Целью исследования является повышение темпа изучения совокупности высказываний в ходе обработки результатов общественного обсуждения законопроектов.

Методы исследований базируются на использовании аппарата математической статистики, теории оптимизации, моделей представления знаний, моделей семантической фильтрации текстов, методов классификации, кластеризации и реферирования текстов.

Научная новизна работы обусловлена следующим:

1. Для решения задачи построения модели гиперкуба использованы зависимости между атрибутами комментариев.

2. Для решения задачи выделения типичных комментариев в ячейках гиперкуба применены кластеризация на основе матриц подобия и принцип наибольшего содержательного подобия со сводным рефератом кластера.

3. Решение задачи отбора полезных комментариев в ячейках гиперкуба осуществлено на основе максимизации функционала обобщенной полезности комментария, сформированного с учетом выявленной аналитической зависимости свертки атрибутов конструктивных предложений, составляющих комментарий, и предпочтений пользователей от ранга комментария.

Теоретическая ценность диссертации заключается в разработке решений, направленных на развитие моделей представления и обработки знаний в форме гиперкуба для обработки текстовой информации на основе операций определения типичных и полезных текстов.

Практическая значимость исследования определяется возможностью повышения оперативности анализа мнений интернет-пользователей по обсуждаемым вопросам на основе технической реализации и применения разработанных модели и алгоритмов в деятельности информационно-аналитических и ситуационных центров государственных органов.

Реализация результатов исследования осуществлена в практику деятельности ведомственных аналитических подразделений и в учебный процесс Академии ФСО России, что подтверждается соответствующими актами.

Область исследования. Содержание диссертации соответствует паспорту специальности 05.13.17 – Теоретические основы информатики (технические науки) по следующим областям исследований:

п. 1. Исследование, в том числе с помощью средств вычислительной техники, информационных процессов, информационных потребностей коллективных и индивидуальных пользователей;

п. 2. Исследование информационных структур, разработка и анализ моделей информационных процессов и структур;

п. 5. Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечения, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.

Научные результаты, выносимые на защиту:

1. Модель оперативной аналитической обработки комментариев, отличающаяся от известных способом формирования гиперкуба на основе многомерного представления мнения, а также набором операций погружения в текстовые данные, учитывающих неравномерность распределения высказываний в комментариях интернет-пользователей.

2. Комплекс алгоритмов, включающий алгоритм формирования выборки типичных комментариев, учитывающий содержательное подобие комментариев и сходство с типичным представителем мнения для приоритетного отбора комментариев, и алгоритм формирования выборки полезных комментариев, учитывающий атрибуты конструктивных предложений и предпочтения интернет-пользователей для приоритетного отбора комментариев.

3. Научно-технические предложения по построению и применению системы оперативной аналитической обработки комментариев, позволяющие производить их обработку с учетом неравномерного распределения высказываний и сходства с типичным представителем мнения при заданном бюджете времени.

Достоверность выводов и рекомендаций обусловлена корректностью применения методов математической статистики, теории оптимизации, методов обработки текстов, воспроизводимостью и проверяемостью теоретических и экспериментальных результатов, согласованностью с практикой, внутренней непротиворечивостью, практической реализацией полученных результатов при разработке автоматизированных способов анализа комментариев.

Личный вклад соискателя. Все изложенные в диссертации результаты исследования получены соискателем лично с учетом замечаний и рекомендаций научного руководителя.

Апробация результатов диссертационного исследования.

Результаты исследования обсуждались на Международной молодежной научно-практической конференции "ИНФОКОМ-2013" (г. Ростов-на-Дону, Северо-Кавказский филиал Московского технического университета связи и информатики, 2013), VIII Всероссийской межведомственной научной конференции "Актуальные проблемы развития технологических систем государственной охраны, специальной связи и специального информационного обеспечения" (г. Орёл, Академия ФСО России, 2013), V Межвузовской научно-практической конференции "Проблемы и перспективы совершенствования охраны Государственной границы" (г. Калининград, Калининградский пограничный институт ФСБ России, 2012), VII Научно-практической конференции "Проблемы развития технологических систем государственной охраны, специальной связи и специального информационного обеспечения" (г. Орёл, Академия ФСО России, 2011).

Публикации. По теме диссертации опубликовано 8 печатных работ, из них 3 – в журналах из Перечня ВАК Минобрнауки РФ, получено 2 свидетельства о регистрации программ для ЭВМ в Роспатенте.

Объем и структура работы. Диссертация состоит из введения, четырех разделов, заключения и приложений. Работа изложена на 207 страницах машинного текста, включая 79 рисунков, 69 таблиц, список использованных источников из 117 наименований и 4 приложения на 40 страницах.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследования, его научная новизна. Определены объект, предмет, цель исследования. Перечислены научные результаты, выносимые на защиту, приведены сведения о публикациях, апробациях и реализации результатов.

В первом разделе проведен анализ процесса обработки комментариев, произведен анализ решений, позволяющих повысить эффективность указанного процесса с учетом многомерности мнений, и произведена постановка задачи.

Показано, что обработка комментариев осуществляется комиссией экспертов и связана со значительными трудозатратами по изучению и обобщению содержащихся в них высказываний. Выявлено, что в процессе анализа комментариев и формирования аналитического обзора важную роль играют классификация комментариев, анализ интенсивности мнений по классам, установление типичных мнений и подкрепление их полезными комментариями.

Анализ источников, характеризующих понятие "мнение", позволил задать его теоретико-множественное представление в виде кортежа:

$$M = \langle W, F, V, O, \{A_i\}, D, C, S, T, Q, E, U \rangle, \quad (1)$$

где W – текст высказывания; F – факт, относительно которого высказывают мнение; V – оценка факта; O – отношение к факту; $\{A_i\}$ – автор(ы) мнения; D – степень доверия мнению (обоснованность мнения); C – цель высказывания мнения; S – оценка стиля высказывания; T – время; Q – аспект факта, по которому высказывают мнение; E – интенсивность проявления мнения; U – оценка полезности мнения.

Установлено, что для обработки текстов, допускающих многомерное представление, предложено использовать модели оперативной аналитической обработки текстов. Выявлено, что существующие модели указанного типа не универсальны и не применимы для эффективного решения задачи обработки комментариев.

В целях формального описания решаемой задачи массив комментариев представлен как двудольный граф $G = (\{\{SENT_v\}, \{COM_v\}\}; \{(sent_{v,m_v}, com_{v,j_v})\})$, где множества

комментариев COM и высказываний $SENT$ поделены на множество классов $\Omega = \bigcup_{v=1}^w \omega_v$

и в каждом классе ω_v существуют: $SENT_v = \bigcup_{m_v=1}^{n_v} sent_{v,m_v}$ – множество высказываний;

$COM_v = \bigcup_{j_v=1}^{k_v} com_{v,j_v}$ – множество комментариев; $L_v = \{l_{v,j_v}\}$ – множество весов коммен-

тариев, определяемых количеством голосов респондентов, отданных за комментарий;

$\Psi_v = \bigcup_{j_v=1}^{k_v} \psi_{v,j_v}$, где $\psi_{v,j_v} : SENT_v \rightarrow COM_v$ – множество функций синтеза комментариев

респондентами (рисунок 1).

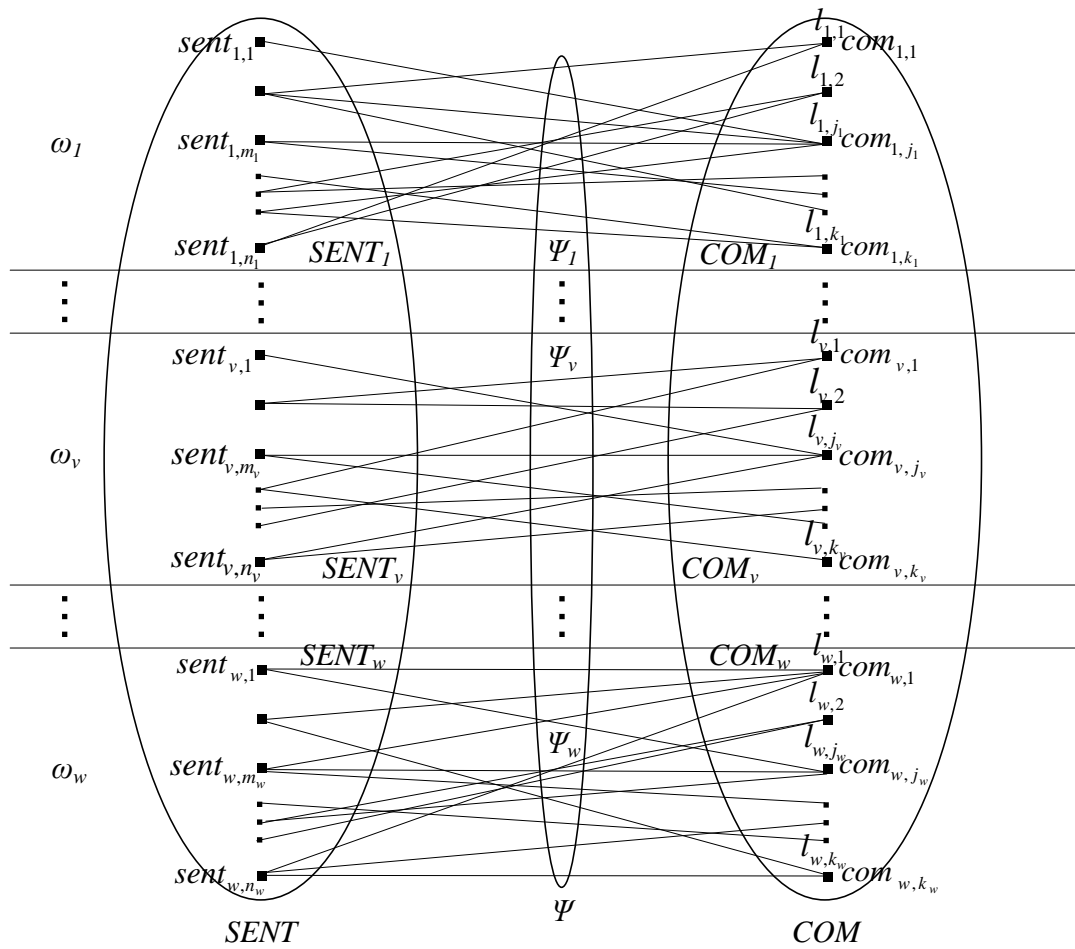


Рисунок 1 – Граф результата обсуждения законопроекта

Требуется: для графа G предложить модель обработки элементов множеств $\{COM_v\}$ при заданном директивном времени обработки Td на основе определения подмножеств $\{com_{v,j_v}\} \subseteq \{COM_v\}$, обеспечивающую выполнение требования, представленного выражением:

$$\begin{cases} \frac{\Delta R_{\{com_{v,j_v}\}}}{\Delta t} \geq \frac{\Delta R_{\{COM_v\}}}{\Delta t}; \\ 0 \leq t \leq Td, \end{cases} \quad (2)$$

где $R_{\{com_{v,j_v}\}} = \frac{|SENT_{\{com_{v,j_v}\}}|}{|SENT_v|}$ – полнота присутствия высказываний в отобранных комментариях множества $\{com_{v,j_v}\}$ в сравнении с комментариями множества COM_v ,

$R_{\{COM_v\}} = \frac{|SENT_{\{COM_v\}}|}{|SENT_v|}$ – полнота присутствия высказываний в последовательно взятых за бюджет времени Td комментариях исходного множества COM_v , t – время.

Учитывая многомерность теоретико-множественного представления мнений, обработку комментариев предложено осуществлять с применением оперативной аналитической обработки текстов на основе гиперкубовой модели вида $hc : a_1 \times \dots \times a_s \rightarrow a_r$, $s, r \in [1, z]$, $s \neq r$, дополненной операциями детализации $f_{TCS} : COM \rightarrow TCS$ и $f_{UCS} : COM \rightarrow UCS$ для обеспечения качественного погружения в массив текстовых комментариев, где множество $TCS = \bigcup_{v=1}^w TCS_v$, $TCS_v = \{com_{v,j_v}\}$, $TCS \subseteq COM$, представляет

собой выборку типичных комментариев, формируемую в результате решения задачи:

$$\begin{cases} |TCS| \rightarrow \min; \\ \sum_{v=1}^w \sum_{j_v=1}^{g_v} y_{v,j_v} \rightarrow \max; \\ Td = const, \end{cases} \quad (3)$$

где $g_v = |TCS_v|$ – мощность множества TCS_v , $y_{v,m_v} = d(sent_{v,m_v})$ – количество комментариев, содержащих высказывание $sent_{v,m_v}$, определяемое степенью соответствующей вершины графа G , характеризующее типичность высказывания $sent_{v,m_v}$, а множество

$UCS = \bigcup_{v=1}^w UCS_v$, $UCS_v = \{com_{v,j_v}\}$, $UCS \subseteq COM$, представляет собой выборку полезных комментариев, формируемую в результате решения задачи:

$$\begin{cases} |UCS| \rightarrow \min; \\ \sum_{v=1}^w \sum_{j_v=1}^{h_v} u_{v,j_v} \rightarrow \max; \\ Td = const, \end{cases} \quad (4)$$

где $h_v = |UCS_v|$ – мощность множества UCS_v , $u_{v,j_v} = f(p_{v,j_v}, l_{v,j_v})$ – свертка показателей p_{v,j_v} и l_{v,j_v} , характеризующая полезность комментария com_{v,j_v} , $p_{v,j_v} = d(com_{v,j_v})$ – количество высказываний в комментарии com_{v,j_v} , определяемое степенью соответствующей вершины графа G . Данная технология представлена на рисунке 2.

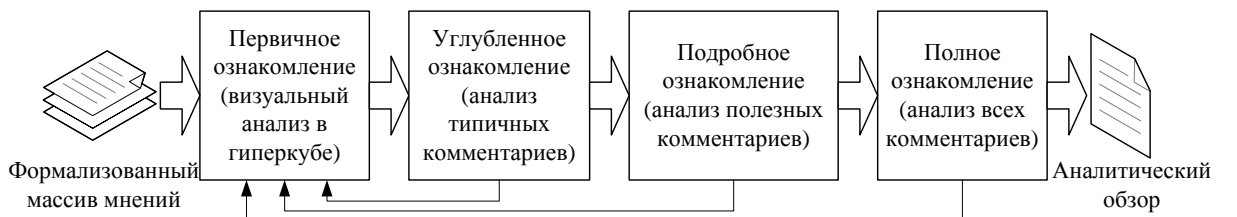


Рисунок 2 – Технология оперативной аналитической обработки комментариев

Во втором разделе приведено описание модели оперативной аналитической обработки комментариев и алгоритмов формирования выборок типичных и полезных комментариев.

Построение гиперкубовой модели осуществляется на основе упорядоченного множества атрибутов A :

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}, \quad (5)$$

где a_1 – глава законопроекта; a_2 – статья законопроекта, к которой относится комментарий; a_3 – идентификатор комментария; a_4 – функциональный стиль комментария; a_5 – оценочная тональность комментария; a_6 – эмоциональная тональность комментария; a_7 – количество комментариев.

На данном множестве экспертным способом устанавливаются функциональные зависимости:

$$DEP = \{a_1 \rightarrow a_2, a_2 \rightarrow a_3, a_3 \rightarrow a_4, a_3 \rightarrow a_5, a_3 \rightarrow a_6, a_2 a_4 a_5 a_6 \rightarrow a_7\}. \quad (6)$$

Полученные зависимости позволяют сформировать измерения $\{D_i\}$:

$$D_1 = \{a_1 \{a_2 \{a_3\}\}\}; \quad (7)$$

$$D_2 = \{a_4\}; \quad (8)$$

$$D_3 = \{a_5\}; \quad (9)$$

$$D_4 = \{a_6\}. \quad (10)$$

В итоге гиперкубовая модель представляется выражением:

$$hc : \{a_1 \{a_2 \{a_3\}\}\} \times \{a_4\} \times \{a_5\} \times \{a_6\} \rightarrow a_7. \quad (11)$$

Решение задачи (3) предлагается осуществлять методом кластеризации комментариев в классах на основе матриц различия, сформированных с использованием матриц подобия с выбором в полученных кластерах типичного комментария.

Для определения квот в классах используется отношение выделенного директивно-го времени к максимальному:

$$Q_v = Q_{v_{\max}} \cdot \frac{Td}{Td_{\max}}. \quad (12)$$

Полученные квоты определяют число кластеров $i = \overline{1, Q_v}$, на которое производится разбиение множества комментариев класса COM_v :

$$cluster : COM_v \rightarrow \{COM_{v_i}\}, \quad (13)$$

где $cluster$ – операция кластеризации, осуществляемая на основе матрицы различия, полученной из матрицы подобия комментариев класса.

В каждом кластере COM_{v_i} производится операция сводного реферирования $svrefer$:

$$\forall COM_{v_i} \ svrefer : COM_{v_i} \rightarrow ref_{v_i}, \quad (14)$$

где ref_{v_i} – сводный реферат комментариев кластера COM_{v_i} , причем режим реферирования должен обеспечивать наилучшее среднее подобие в кластере комментариев. Определение типичного комментария в каждом кластере COM_{v_i} производится посредством операции $typcom$ нахождения комментария $com_{v_{ik}}$, обладающего наилучшим подобием $dist$ с полученным сводным рефератом кластера ref_{v_i} :

$$\forall COM_{v_i} \ typcom : COM_{v_i} \rightarrow com_{v_{ik}} \mid k = \arg \min dist(com_{v_{ik}}, ref_{v_i}). \quad (15)$$

Ознакомление эксперта с выборкой типичных комментариев в классе v осуществляется в пределах квоты Q_v в порядке включения комментариев в выборку с увеличением бюджета времени и ростом числа кластеров COM_{v_i} .

Решение задачи (4) осуществляется путем нахождения максимума функционала обобщенной полезности $f(Q)$ с использованием метода множителей Лагранжа при условии $Td = const$:

$$Q = \arg \max_{Td=const} f(Q), \quad (16)$$

где Q – вектор квот для классов комментариев, оказывающих значимое влияние на обобщенный показатель полезности u_{v,j_v} :

$$u_{v,j_v} = \alpha_p \cdot \frac{p_{v,j_v}}{p_{\max}} + \alpha_l \cdot \frac{l_{v,j_v}}{l_{\max}}, \quad (17)$$

где α_p, α_l – коэффициенты значимости, подбираемые с учетом распределений показателей свертки. Поскольку кривая, характеризующая убывание обобщенного показателя в классах, со средней достоверностью 0,9 аппроксимируется логарифмической функцией вида $g_v(Q_v) = A_v - B_v \cdot \ln(Q_v)$, функционал $f(Q)$ принимает вид:

$$f(Q) = \sum_{v=1}^w f(Q_v); \quad (18)$$

$$f(Q_v) = \int_{Q_v} A_v - B_v \cdot \ln(Q_v) dQ_v; \quad (19)$$

$$f(Q) = \sum_{v=1}^w [Q_v \cdot (A_v + B_v) - B_v \cdot Q_v \cdot \ln(Q_v)]. \quad (20)$$

Задание директивного времени Td осуществляется при условии:

$$0 < Td \leq Td_{\max}, \quad (21)$$

причем общее уравнение бюджетного ограничения по времени примет вид:

$$Td = \sum_{v=1}^w L_v \cdot T1 \cdot Q_v, \quad (22)$$

где L_v – среднее количество слов в комментарии, $T1$ – среднее время чтения одного слова, Q_v – число комментариев выборки класса v , соответственно:

$$Td_{\max} = \sum_{v=1}^w L_v \cdot T1 \cdot Q_{v_{\max}}. \quad (23)$$

В реальных условиях естественное ограничение накладывается и на количество комментариев в выборках классов Q_v :

$$1 \leq Q_v \leq Q_{v_{\max}}. \quad (24)$$

В итоге задача (16) сводится к системе:

$$\left\{ \begin{array}{l} Td = \sum_{v=1}^w L_v \cdot T1 \cdot Q_v; \\ \frac{A_1 - B_1 \cdot \ln(Q_1)}{L_1 \cdot T1} = \frac{A_2 - B_2 \cdot \ln(Q_2)}{L_2 \cdot T1}; \\ \frac{A_2 - B_2 \cdot \ln(Q_2)}{L_2 \cdot T1} = \frac{A_3 - B_3 \cdot \ln(Q_3)}{L_3 \cdot T1}; \\ \dots \\ \frac{A_{w-1} - B_{w-1} \cdot \ln(Q_{w-1})}{L_{w-1} \cdot T1} = \frac{A_w - B_w \cdot \ln(Q_w)}{L_w \cdot T1}; \\ 1 \leq Q_1 \leq Q_{1_{\max}}; \\ \dots \\ 1 \leq Q_w \leq Q_{w_{\max}}; \end{array} \right. \quad (25)$$

решение которой осуществляется численными методами с использованием математических пакетов прикладных программ.

Ознакомление эксперта с выборкой полезных комментариев в классе V осуществляется в пределах квоты Q_V в порядке убывания обобщенного показателя полезности комментария u_{v,j_v} .

Операции погружения на основе предложенной модели осуществляются с использованием алгоритмов формирования выборок типичных и полезных комментариев.

Алгоритм формирования выборки типичных комментариев (решения задачи (3)) представлен на рисунке 3. Сущность данного алгоритма заключается в разбиении, в соответствии с заданным бюджетом времени, исходного множества комментариев в классах (ячейках гиперкуба) на кластеры на основе матриц подобия и определении множества комментариев, по одному в каждом кластере, обладающих наилучшим сходством с типичными представителями кластеров комментариев, в качестве которых выступают сводные рефераты кластеров.

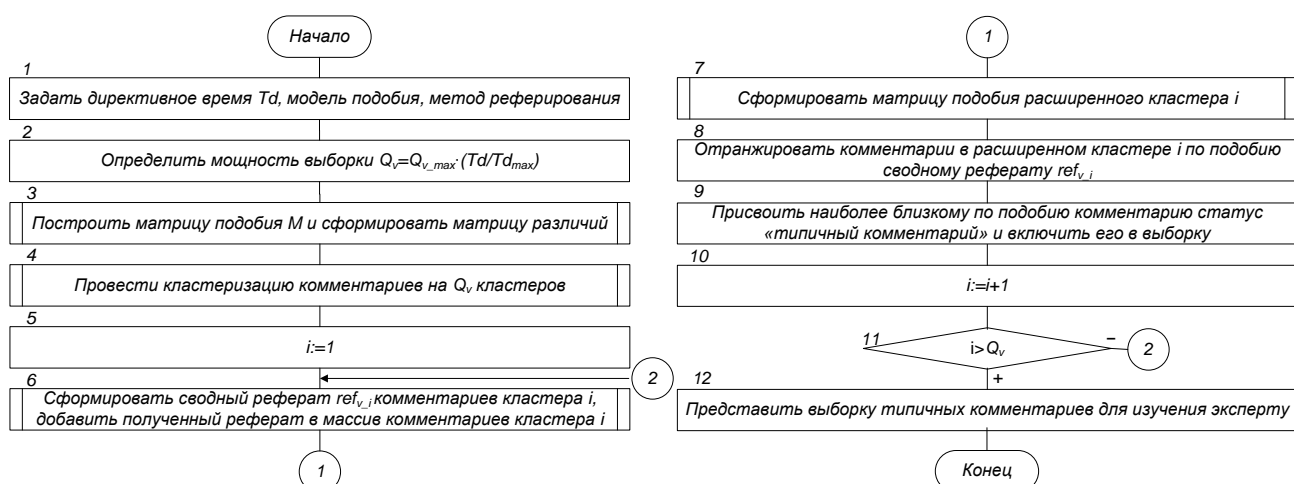


Рисунок 3 – Алгоритм формирования выборки типичных комментариев

В ходе экспериментов установлено, что сводные рефераты обладают наилучшим средним подобием в классе комментариев (таблица 1). Указанное свойство было положено в основу процедуры определения типичных комментариев.

Таблица 1 – Свойство наилучшего среднего подобия сводных рефератов комментариев различных классов на примере законопроекта "Об образовании" (модель подобия – марковская, рефераты получены методом сканирующих статистик)

Текст (класс "34-ГП-Н-Ми")	Ср. коэф. подобия	Текст (класс "11-ОД-Н-Ми")	Ср. коэф. подобия	Текст (класс "1-ОД-Н-Ми")	Ср. коэф. подобия
edu 34-3.txt	0,626469	edu 11-4.txt	0,163084	edu 1-3.txt	0,324475
edu 34-4.txt	0,62749	edu 11-1.txt	0,163153	edu 1-1.txt	0,326142
edu 34-1.txt	0,629656	edu 11-2.txt	0,163153	edu 1-2.txt	0,3335
4166	0,986917	2343	0,226228	7732	0,600645
520	1,071068	2823	0,289161	5520	0,658391
458	1,078332	2346	0,306828	6276	0,667973
..

На шаге 1 алгоритма задается бюджет времени Td с учетом ограничения $0 \leq Td \leq Td_{\max}$, производится выбор модели подобия и метода реферирования..

На шаге 2, исходя из заданного бюджета времени и мощности класса $Q_{v_{\max}} = |COM_v|$, определяется мощность выборки типичных комментариев.

На шаге 3 для выбранного класса комментариев строится матрица подобия M_v , на основе которой, в зависимости от используемой модели подобия, формируется матрица различий (таблица 2).

Таблица 2 – Фрагмент матрицы различий (модель подобия – марковская, класс "7-42-Н-В-Ми" законопроекта "О полиции")

Ид. ком.	19074	359	1658	1779	2715	2839	3875	...
19074	0	2,7175	3,0272	3,0839	3,1179	2,5739	4,2059	...
359	2,6959	0	0,9163	1,0651	0,3536	0,156	0,8233	...
1658	2,9231	0,9522	0	1,1987	1,3682	0,1889	2,7109	...
1779	2,8662	1,0529	1,206	0	0,2912	0,926	2,8638	...
2715	3,0401	0,3536	1,384	0,2912	0	1,1362	3,1515	...
2839	2,6374	0,156	0,1889	0,901	1,1147	0	0,4452	...
3875	4,1056	0,8233	2,5657	2,6757	2,8837	0,4452	0	...
...

На шаге 4 к полученной матрице различий применяется метод иерархического объединения для кластеризации комментариев на требуемое количество кластеров Q_v в соответствии с заданным бюджетом времени Td (рисунок 4).

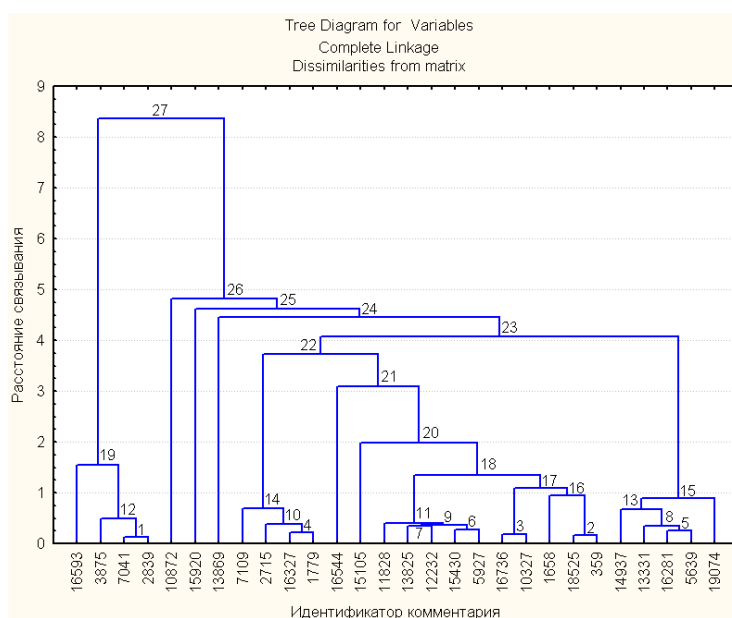


Рисунок 4 – Результат кластеризации комментариев (метод кластеризации – иерархическое объединение, класс "7-42-Н-В-Ми" законопроекта "О полиции")

На шаге 6 в полученных кластерах строятся сводные рефераты комментариев и включаются в массив соответствующего кластера.

На шаге 7 формируются матрицы подобия в расширенных полученных сводными рефератами кластерах.

На шаге 8 определяются комментарии, обладающие наилучшим сходством со сводным рефератом своего кластера (таблица 3).

Таблица 3 – Выбор типичного комментария в 15 кластере (модель подобия – марковская, класс "7-42-Н-В-Ми" законопроекта "О полиции")

Текст	5639.txt	13331.txt	14937.txt	16281.txt	19074.txt	all15ref3.txt
all15ref3.txt	0,2838	0,0591	2,7037	0,0834	3,1937	0
13331.txt	0,2624	0	0,6804	0,3556	0,8036	0,0591
16281.txt	0,2474	0,3556	0,674	0	0,8227	0,0834
5639.txt	0	0,2624	0,6736	0,2474	0,8242	0,2838
14937.txt	0,6736	0,6804	0	0,674	0,8996	2,5486

Шаги 6–10 повторяются для каждого кластера, полученного на шаге 4.

На шаге 12 совокупность отобранных в каждом из кластеров комментариев составляет выборку типичных комментариев указанного экспертом класса, которые подлежат изучению в порядке включения в выборку при увеличении числа кластеров (таблица 4).

Таблица 4 – Формирование выборки типичных комментариев при увеличении числа кластеров (модель подобия – марковская, класс "7-42-Н-В-Ми" законопроекта "О полиции")

К-во комм.	Номер кластера																											Поряд. включ.	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27		
1	2839	359	10327	1779	5639	5927	12232	5639	5927	1779	5927	2839	5639	1779	19074	359	359	359	2839	359	359	19074	19074	19074	19074	19074	19074	15	
2	7041	18525	16736	16327	16281	15430	13825	16281	15430	16327	15430	7041	16281	16327	5639	18525	18525	18525	7041	18525	18525	18525	5639	5639	5639	5639	5639	24	
3																													23
4																													6
5																													17
6																													27
7																													14
8																													12
9																													26
10																													13
11																													21
12																													7
13																													19
14																													22
15																													11
16																													9
17																													8
18																													2
19																													25
20																													20
21																													16
22																													5
23																													4
24																													3
25																													1
26																													28
27																													18
28																													10

Алгоритм формирования выборки полезных комментариев (решения задачи (4)) представлен на рисунке 5. Сущность данного алгоритма заключается в определении обобщенного показателя полезности комментариев, выявлении значимых атрибутов комментариев, установлении параметров аналитических функций предельной полезности комментариев в классах, образованных значимыми атрибутами, определении стоимостных параметров, задании бюджета времени и решении оптимизационной задачи с использованием метода множителей Лагранжа.

Применение метода множителей Лагранжа обусловлено возможностью построения аналитических выражений для убывающих функций предельной полезности комментариев в каждом классе, наличием временной стоимости обработки комментариев, связанной с темпом чтения и длиной комментария, и возможностью сформировать уравнение бюджетного ограничения (в данной задаче стоимость измеряется в единицах времени).

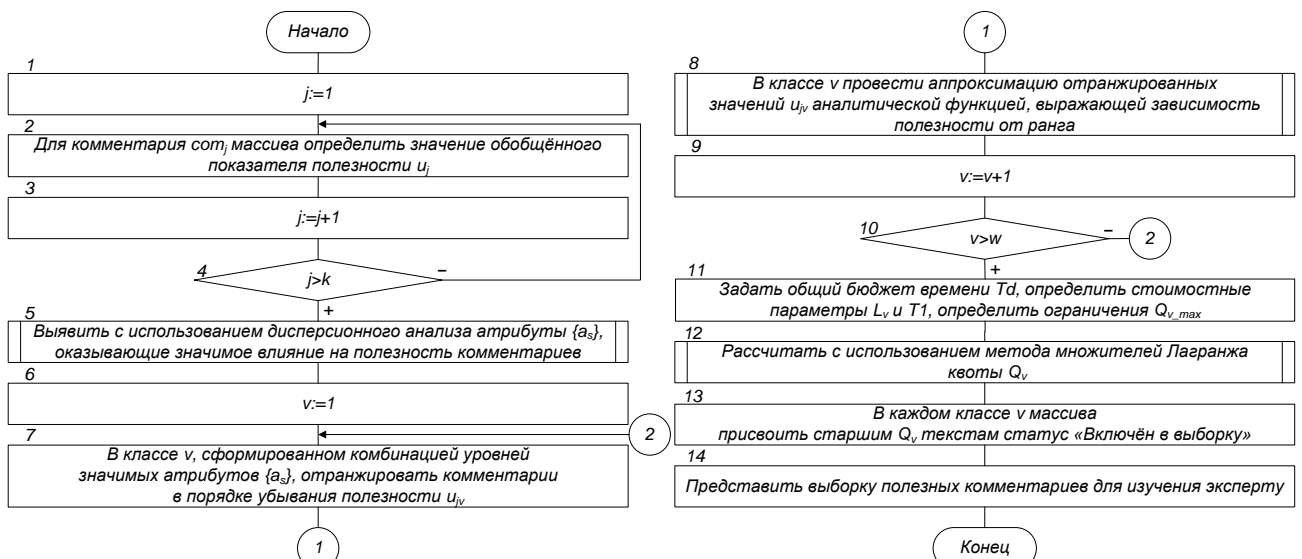


Рисунок 5 – Алгоритм формирования выборки полезных комментариев

На шаге 2 для определения обобщенного показателя полезности применяется аддитивная свертка числа голосов, отданных за комментарий, и числа конструктивных предложений, с учетом распределений данных параметров, фрагмент расчета представлен в таблице 5.

Таблица 5 – Результаты расчета обобщенного показателя полезности комментариев к законопроекту "О полиции"

Идентификатор комментария	Голосов ("лайков")	Конструктивных предложений	Голосов нормир.	Предложений нормир.	Свертка
...
11099	1	3	0,00108	0,157895	0,013468
11100	1	1	0,00108	0,052632	0,005152
11101	0	4	0	0,210526	0,016632
11102	0	0	0	0	0
11103	0	1	0	0,052632	0,004158
...

На шаге 5 атрибуты, оказывающие значимое влияние на обобщенный показатель полезности, определяются с использованием дисперсионного анализа (таблица 6).

Таблица 6 – Результаты дисперсионного анализа для комментариев к законопроекту "О полиции"

Эффект	SS	Степ. свободы	MS	F	p
Статья	0,064	56	0,001	2,75	0,000*
Стиль	0,15	4	0,037	89,6	0,000*
Тон(эмоц)	0,001	2	0	0,77	0,462
Тон(оцен)	0,003	1	0,003	6,56	0,010*

Шаги 7 и 8 необходимы для получения аналитических выражений функций предельной полезности. На шаге 7 в классах COM_v , сформированных комбинацией выявленных на шаге 5 значимых атрибутов $\{a_s\}$, производится ранжирование комментариев $\{com_{v,j_v}\}$ по убыванию обобщенного показателя полезности u_{v,j_v} , полученного для каждого комментария на шаге 2.

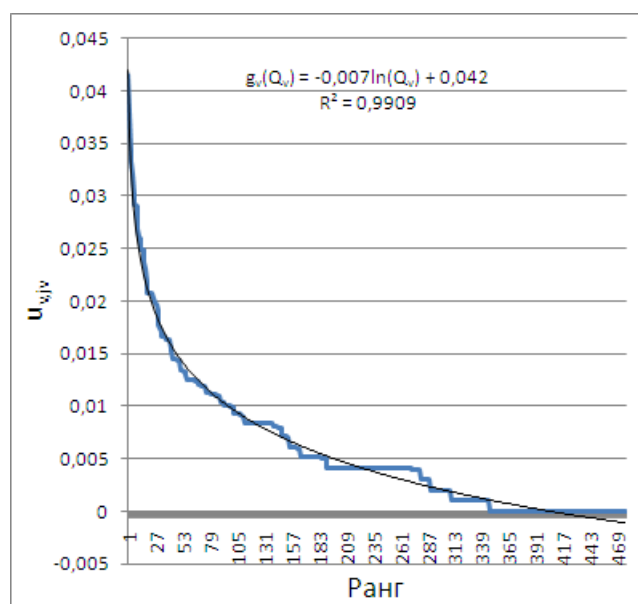


Рисунок 6 – Пример аппроксимации графика предельной полезности комментариев класса "2-Науч-Минор" к законопроекту "О полиции" логарифмической кривой с параметрами и достоверностью аппроксимации

На шаге 8 полученные кривые аппроксимируются логарифмической функцией $g_v(Q_v)$ (рисунок 6). Результатом шага 8 для всех классов является совокупность параметров всех функций предельной полезности.

На шаге 11 определяются стоимостные параметры и ограничения, характеризующие массив комментариев: максимальное число комментариев Q_{v_max} в классах, сформированных комбинацией значимых атрибутов $\{a_s\}$; средняя длина комментария в классе L_v (слов); задается бюджет времени Td . Результаты выполнения шагов 8 и 11 представлены в таблице 7.

Таблица 7 – Фрагмент таблицы с параметрами функций предельной полезности и ограничениями для классов комментариев COM_v к законопроекту "О полиции"

v	Статья	Стиль	Тон.(оцен.)	A_v	$B_v (-B_v \cdot \ln Q)$	Достов. аппр.	Q_{v_max}	L_v
...
132	18	газет.-публ.	минорная	0,0303	0,009	0,9234	28	71,32
133	19	газет.-публ.	минорная	0,0176	0,005	0,9626	36	49,94
134	20	газет.-публ.	минорная	0,0221	0,006	0,9284	25	49,08
135	21	газет.-публ.	минорная	0,023	0,008	0,9344	20	60,7
...

Совокупность параметров, полученных в результате выполнения шагов 8 и 11, используется на шаге 12 в решении системы (26), здесь представлен фрагмент системы уравнений при $Td = 0,5 \cdot Td_{max}$ с неизвестными параметрами выборки полезных комментариев к законопроекту "О полиции":

$$\left\{ \begin{array}{l}
 0,5 \cdot 348800 = \sum_{v=1}^{472} L_v \cdot 0,3 \cdot Q_v; \\
 \frac{0,076 - 0,013 \cdot \ln(Q_1)}{97,77 \cdot 0,3} = \frac{0,042 - 0,007 \cdot \ln(Q_2)}{73,45 \cdot 0,3}; \\
 \frac{0,042 - 0,007 \cdot \ln(Q_2)}{73,45 \cdot 0,3} = \frac{0,051 - 0,009 \cdot \ln(Q_3)}{84,82 \cdot 0,3}; \\
 \dots \\
 \frac{0,003 - 0,004 \cdot \ln(Q_{471})}{10 \cdot 0,3} = \frac{0,001 - 0,001 \cdot \ln(Q_{472})}{15 \cdot 0,3}; \\
 1 \leq Q_1 \leq 496; \\
 \dots \\
 1 \leq Q_{472} \leq 1.
 \end{array} \right. \quad (26)$$

На шаге 13 полученные на шаге 12 квоты применяются к полученным на шаге 7 ранжировкам, вошедшие в квоту комментарии выбираемого аналитиком класса составляют выборку полезных комментариев.

Ознакомление с выборкой полезных комментариев класса COM_v производится в пределах квоты Q_v в порядке убывания обобщенного показателя полезности u_{v,j_v} .

В третьем разделе произведено оценивание адекватности модели и эффективности разработанных алгоритмов. Для этого были исследованы статьи законопроектов "О полиции" и "Об образовании". В ходе исследования было установлено, что статьи законопро-

ектов содержат различное количество пунктов и подпунктов, что обуславливает большее либо меньшее разнообразие высказываний. При этом распределение статей по количеству объектов комментирования подчинено экспоненциальному закону.

Для экспериментов были отобраны классы комментариев к статьям, содержащим наибольшее, наименьшее и среднее количество объектов комментирования, т.е. по три класса для двух законопроектов. В указанных классах был произведен экспертный анализ комментариев, выделены высказывания (таблица 8) и составлены распределения высказываний по комментариям классов (таблица 9).

Таблица 8 – Фрагмент таблицы с высказываниями класса "34-Науч.-Низ.-Минор." к законопроекту "О полиции"

m_v	Содержание высказывания $sent_{v,m_v}$, $v = "6-34-Н-Н-Ми"$
1	Исключить привлечение к оперативно-розыскной деятельности сотрудников внутренней службы, юстиции и стажеров.
2	Конкретизировать требования (статус, права и обязанности, уровень образования, степень подготовки, экипировка и пр.) к привлекаемым к охране общественного порядка и пр. задачам иных сотрудников органов внутренних дел (ОВД) и стажеров
3	Включить в перечень оснований привлечения сотрудников ОВД объявление режима ЧС, природные и техногенные катастрофы, ликвидацию последствий стихийных действий, проведение контртеррористических операций, проведение массовых мероприятий.
4	Уточнить категории сотрудников ОВД, на которые распространяется действие статьи
5	Предоставить гражданину право удостовериться, что привлеченный сотрудник уполномочен совершать в отношении него действия
...	...

Таблица 9 – Пример распределения высказываний в классе "34-Науч.-Низ.-Минор" к законопроекту "О полиции"

Ид. комм.	Ид. высказывания							
	1	2	3	4	5	6	7	8
2393	+	+						
2505		+	+					
5601				+				
6205		+			+			
6239		+				+		
14322		+						
15793		+						
15907							+	+
17329				+				

Оценивание адекватности модели проводилось путем сравнения типичных высказываний, выделенных и включенных в обзор экспертами, с высказываниями, полученными с использованием модели.

Для оценки адекватности использовалось выражение:

$$r_v = \frac{q_v - k_v}{k_v}, \quad (27)$$

где k_v – мощность класса комментариев ω_v ; q_v – мощность подмножества класса комментариев ω_v (выборки заданного типа, подмножества исходного массива), обеспечива-

ющая охват всех типичных высказываний в классе, что позволило оценить адекватность модели на основе оценки обобщающей способности формируемых выборок исследуемых классов.

В ходе оценивания установлено, что модель обеспечивает выделение высказываний, совпадающих по смыслу с высказываниями, приведенными в обзоре в качестве типичных, при этом выделение типичных высказываний с использованием модели осуществляется в среднем на 20% быстрее по сравнению с выделением этих же высказываний путем последовательного чтения (рисунок 7).

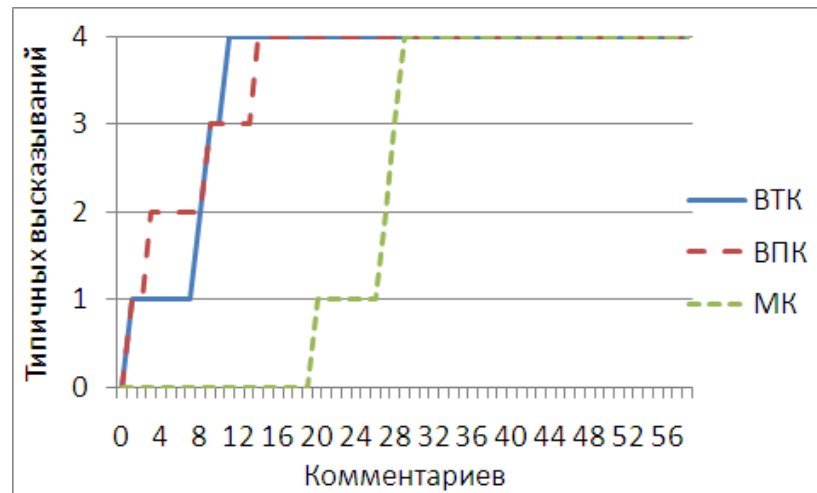


Рисунок 7 – Пример выделения типичных высказываний с использованием выборок типичных и полезных комментариев для класса "3-12" к законопроекту "О полиции"

Для оценивания эффективности алгоритмов были определены множества комментариев, предоставляемые для изучения в соответствии с выделенным бюджетом времени с шагом $\Delta t = 0,05 \cdot Td_{\max}$. Полученные множества пошагово сравнивались с множествами не отранжированных по типичности и полезности комментариев при одинаковом бюджете времени по четырем видам полноты (таблица 10).

Таблица 10 – Виды полноты отбора высказываний

По учету веса высказываний	По идентификации повторов высказываний	
	Без идентификации повторов высказываний	С идентификацией повторов высказываний
Без учета веса высказываний	R1	R3
С учетом веса высказываний	R2	R4

В ходе экспериментов было установлено, что средний прирост полноты зависит от способа ее расчета, в зависимости от учета веса высказываний и идентификации повторов высказываний. Оцениваемый алгоритм формирования выборки типичных комментариев показал наилучшие результаты при оценке по среднему приросту полноты вида R4 (рисунок 8), а алгоритм формирования выборки полезных комментариев – при оценке по среднему приросту полноты вида R1 при $\Delta t = 0,05 \cdot Td_{\max}$ (рисунок 9).

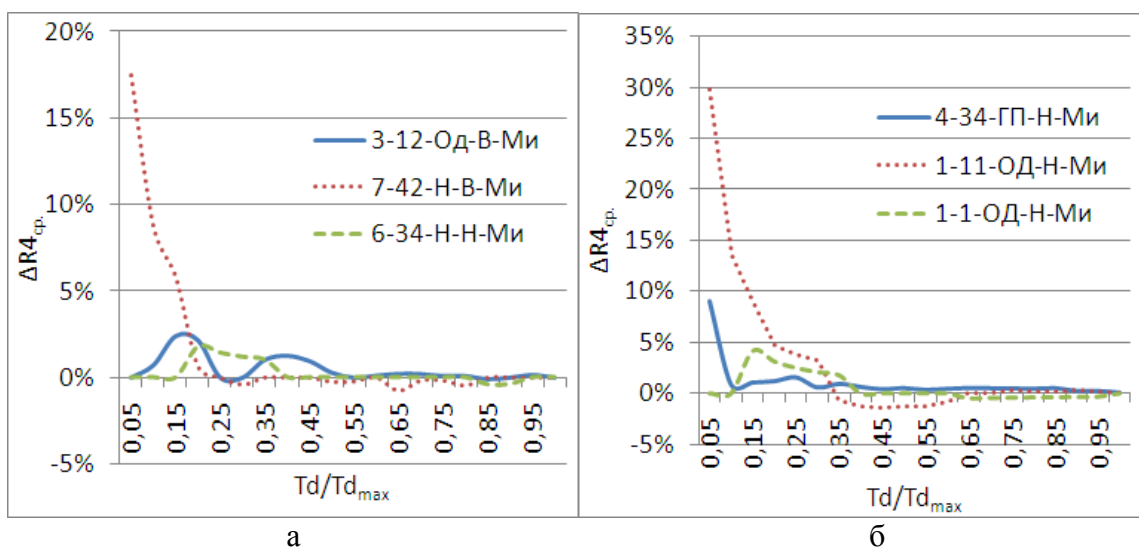


Рисунок 8 – Абсолютный выигрыш в среднем приросте полноты вида R4 при $\Delta t = 0,05 \cdot Td_{\max}$ для выборки типичных комментариев к трем классам с различным числом объектов комментирования законопроекта:
а – "О полиции" б – "Об образовании"

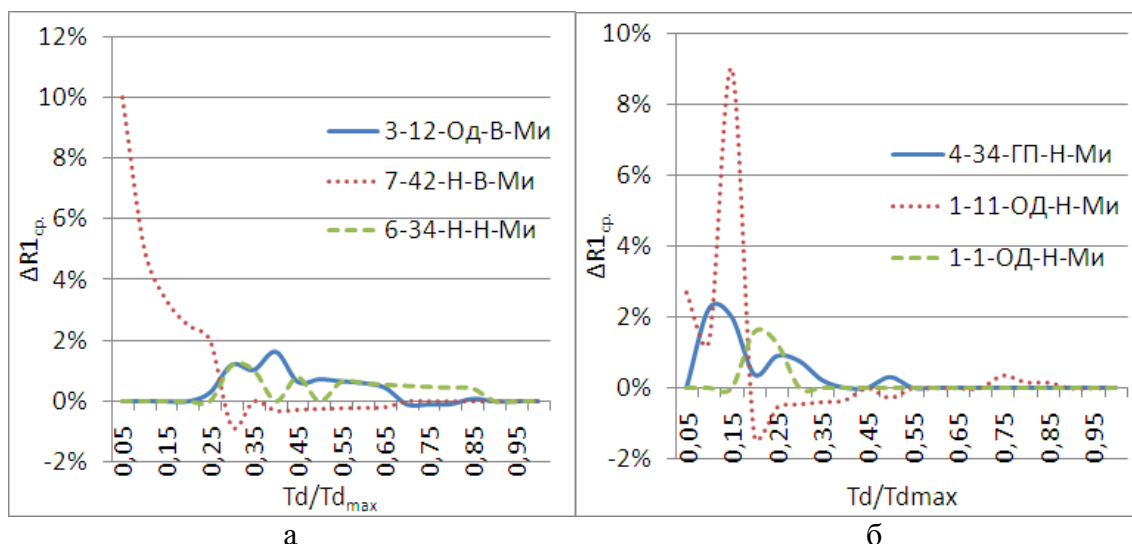


Рисунок 9 – Абсолютный выигрыш в среднем приросте полноты вида R1 при $\Delta t = 0,05 \cdot Td_{\max}$ для выборки полезных комментариев к трем классам с различным числом объектов комментирования законопроекта:
а – "О полиции" б – "Об образовании"

Эксперименты показали, что для различных законопроектов и статей с различным количеством объектов комментирования предложенные алгоритмы позволяют обеспечить опережающий темп изучения совокупности высказываний экспертом, как правило, лишь в первую половину от максимального бюджета времени, что связано с тем, что значительная часть комментариев содержит по одному уникальному высказыванию.

В четвертом разделе описаны структура и методика применения системы оперативной аналитической обработки комментариев. Структура системы представлена на рисунке 10. В состав системы входят следующие подсистемы:

- подсистема идентификации комментариев;
- подсистема классификации комментариев;
- блок стандартного OLAP-анализа;
- подсистема формирования выборки типичных комментариев;
- подсистема формирования выборки полезных комментариев.

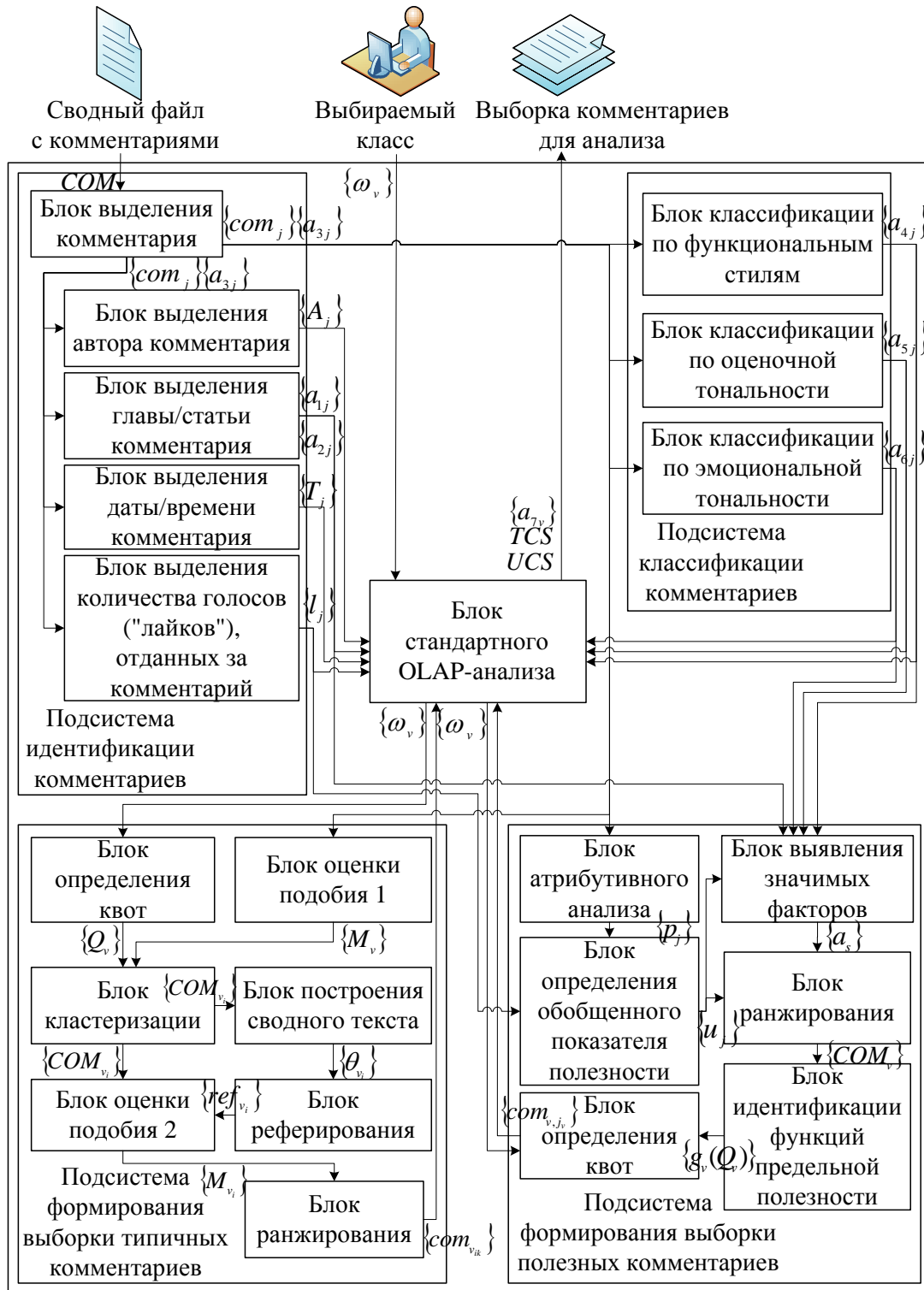


Рисунок 10 – Структура системы оперативной аналитической обработки комментариев

Рассмотрим назначение и функции блоков данных подсистем.

Подсистема идентификации комментариев предназначена для разделения сводно-го файла с комментариями на отдельные файлы комментариев и выявления содержащихся в сводном файле характеристик комментариев ("служебной" информации).

В *блоке выделения комментария* методом лексического анализа осуществляется поиск начала и окончания комментария, отделяется "служебная" информация, производится присвоение комментарию идентификатора и сохранение комментария в отдельный файл.

В блоке выделения автора комментария методом лексического анализа выделяются сведения об авторе (или авторах) комментария и заносятся в таблицу параметров.

В блоке выделения главы/статьи комментария методом лексического анализа выделяются сведения о главе и статье комментария и заносятся в таблицу параметров.

В блоке выделения даты/времени комментария методом лексического анализа выделяются сведения о дате и времени комментария и заносятся в таблицу параметров.

В блоке выделения количества голосов, отданных за комментарий, методом лексического анализа выделяются сведения о количестве голосов в поддержку комментария и заносятся в таблицу параметров.

Подсистема классификации комментариев предназначена для определения оценочных характеристик комментариев с целью разбиения исходного множества комментариев на классы и последующего анализа высказываний в комментариях с одинаковыми значениями классификационных признаков.

В блоках классификации по функциональным стилям, оценочной и эмоциональной тональностям с использованием словарных и статистических методов обработки текстов производится определение функционального стиля, оценочной и эмоциональной тональностей комментария.

Блок стандартного OLAP-анализа предназначен для оперативной аналитической обработки комментариев с использованием сведений об их разбиении и распределении по классам с целью установления очередности обработки в зависимости от интенсивности мнений и предпочтений эксперта.

Подсистема формирования выборки типичных комментариев предназначена для детализации выбираемого класса комментариев на основе решения системы (3) и учитывает неравномерность распределения высказываний в комментариях посредством определения содержательного подобию комментариев и сходства с типичным представителем мнения.

В блоке определения квот с учетом бюджета времени определяется мощность выборки.

В блоке оценки подобию 1 с использованием выбираемой модели подобию строится матрица подобию комментариев класса, на основании которой формируется матрица различий.

В блоке кластеризации полученная матрица различий применяется для кластеризации комментариев на число кластеров, соответствующее мощности выборки.

В блоке построения сводного текста производится объединение комментариев кластера в один файл.

В блоке реферирования с использованием выбранного метода реферирования производится построение сводного реферата. Режим реферирования должен обеспечивать наилучшее среднее подобие полученного сводного реферата на множестве текстов, для которых он строится.

В блоке оценки подобию 2 с использованием модели подобию, аналогичной модели в блоке оценки подобию 1, определяется комментарий кластера, обладающий наилучшим подобием с полученным сводным рефератом кластера.

В блоке ранжирования определяется порядок ознакомления с комментариями выборки методом установления последовательности включения комментариев в выборку с ростом бюджета времени.

Подсистема формирования выборки полезных комментариев предназначена для детализации выбираемого класса комментариев на основе решения системы (4) и учитывает неравномерность распределения высказываний в комментариях посредством определения атрибутов конструктивных предложений и голосов интернет-пользователей.

В блоке атрибутивного анализа методом лексического анализа с применением словарей синонимичных атрибутов трех типов (атрибуты конструктивных предложений

по добавлению, изменению, удалению фрагментов статей законопроекта) производится выявление и подсчет содержащихся в комментарии конструктивных предложений.

В блоке *определения обобщенного показателя полезности* производится расчет данного показателя посредством свертки нормированных значений числа конструктивных предложений в комментарии и числа голосов, отданных за комментарий, с учетом распределений данных параметров.

В блоке *выявления значимых факторов* методом дисперсионного анализа выявляются классификационные признаки, оказывающие значимое влияние на значение обобщенного показателя полезности.

В блоке *ранжирования* для каждой комбинации значимых классификационных признаков производится сортировка комментариев по убыванию значения обобщенного показателя полезности.

В блоке *идентификации функций предельной полезности* производится аппроксимация полученных в блоке ранжирования рядов значений аналитическими (логарифмическими) функциями и фиксация параметров данных функций.

В блоке *определения квот* с учетом общего бюджета времени, темпа чтения, а также функций предельной полезности, средней длины комментария и ограничения на максимальную мощность выборки для каждой комбинации значимых классов, методом множителей Лагранжа определяется мощность выборки для каждой такой комбинации. Знакомление с комментариями выбираемой комбинации классов производится в пределах данной выборки в последовательности, установленной в блоке ранжирования.

Аналитическую обработку комментариев с использованием предложенной системы рекомендуется проводить путем применения следующей методики (рисунок 11).

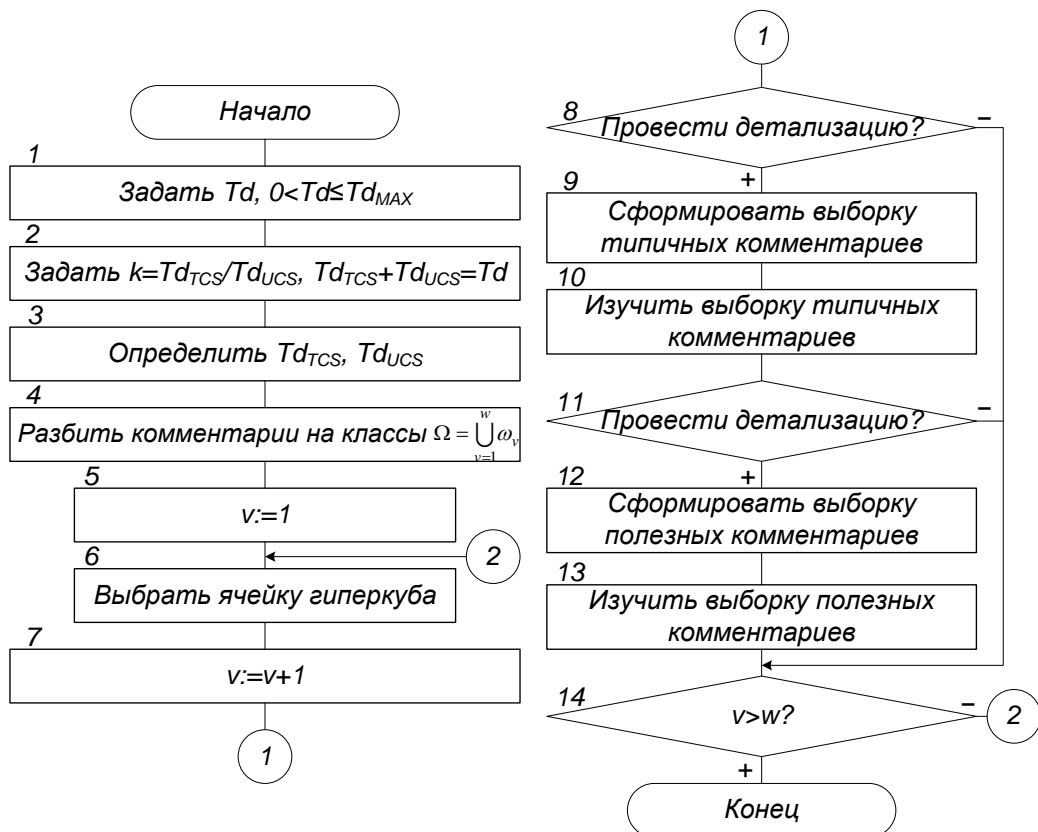


Рисунок 11 – Методика оперативной аналитической обработки комментариев

На шаге 1 задается директивное время на изучение комментариев, оно не должно превышать время, требующееся для изучения всех комментариев.

На шаге 2 производится распределение заданного директивного времени между типичными и полезными комментариями путем задания коэффициента k .

На шаге 3 проводится расчет бюджета времени для изучения типичных и полезных комментариев на основе заданных директивного времени и коэффициента k .

На шаге 4 производится классификация комментариев по стилям, эмоциональной и оценочной тональности, статьям законопроекта.

На шаге 6 выбирается ячейка гиперкуба, т.е. интересующая эксперта комбинация классов.

На шаге 8 принимается решение о проведении детализации с использованием выборки типичных комментариев.

На шаге 9 производится формирование выборки типичных комментариев, алгоритм выполняется каждый раз для выбранной комбинации классов.

На шаге 10 проводится изучение выборки типичных комментариев.

На шаге 11 принимается решение о проведении детализации с использованием выборки полезных комментариев.

На шаге 12 производится формирование выборки полезных комментариев, алгоритм выполняется при первой детализации для всех комбинаций классов, при последующих детализациях производится обращение к результатам работы алгоритма для требуемой комбинации классов.

На шаге 13 проводится изучение выборки полезных комментариев.

Иллюстрация применения методики в ходе аналитической обработки комментариев класса "1-11-Од-Н-Ми" (глава 1, статья 11, стиль – официально-деловой, эмоциональная тональность – низкая, оценочная тональность – минорная) к законопроекту "Об образовании" представлена на рисунке 12. В указанном примере детализация проведена при $Td = 0,35 \cdot Td_{\max}$, $k = 0,75$. При данных параметрах эксперту предоставляются для изучения три типичных комментария (идентификаторы обведены темными овалами) и двенадцать полезных комментариев (идентификаторы обведены светлыми овалами). В данном классе комментариев выявлено три обобщенных мнения, представленных типичными комментариями и подкрепленных одним или несколькими полезными комментариями.

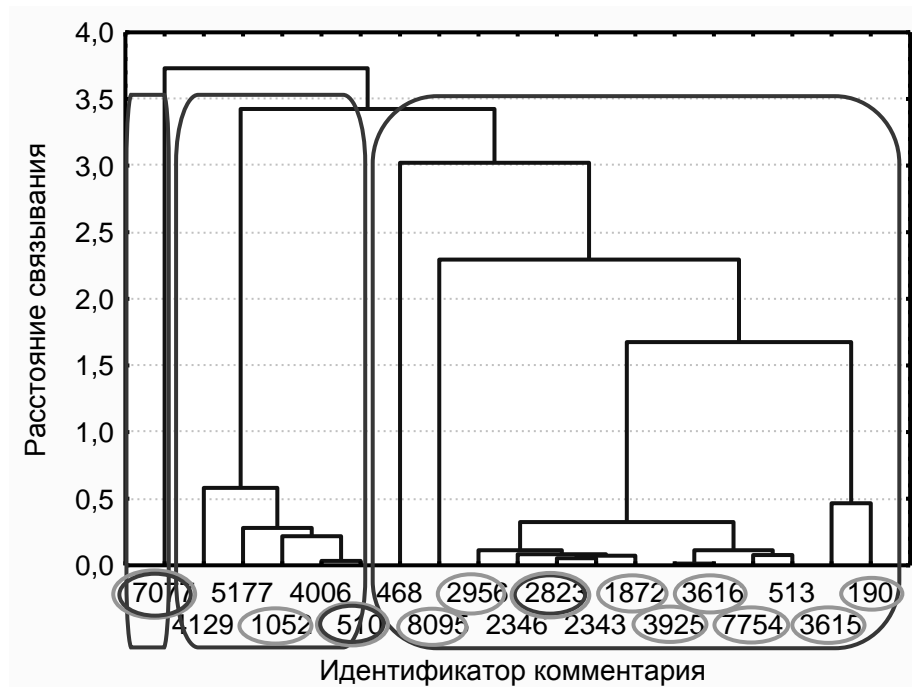


Рисунок 12 – Иллюстрация применения методики

В заключении сформулированы выводы по результатам проведенного исследования.

Основные результаты и выводы работы:

1. Проведен анализ процесса обработки комментариев, полученных в ходе общественного обсуждения законопроектов. Установлено противоречие между потребностью в эффективных моделях обработки комментариев и отсутствием качественной модели обработки комментариев. Представлен вариант разрешения противоречия.

2. Разработана модель оперативной аналитической обработки комментариев, отличающаяся от известных способом формирования гиперкуба на основе многомерного представления мнения, а также набором операций погружения в текстовые данные, учитывающих неравномерность распределения высказываний в комментариях интернет-пользователей.

3. Синтезирован алгоритм формирования выборки типичных комментариев, учитывающий содержательное подобие комментариев и сходство с типичным представителем мнения для приоритетного отбора комментариев.

4. Синтезирован алгоритм формирования выборки полезных комментариев, учитывающий атрибуты конструктивных предложений и предпочтения интернет-пользователей для приоритетного отбора комментариев.

5. Предложены структура и методика применения системы оперативной аналитической обработки комментариев, позволяющие производить их обработку с учетом неравномерного распределения высказываний и сходства с типичным представителем мнения при заданном бюджете времени.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в журналах из Перечня ВАК

1. Толкунов, А. А. Модель оперативной аналитической обработки комментариев с операциями детализации на основе выявления типичных и полезных текстов / А. А. Толкунов // Научное обозрение. – 2013. – № 9. – С. 368–371.

2. Толкунов, А. А. Алгоритмы формирования выборок типичных и полезных комментариев в ходе обработки результатов общественного обсуждения законопроектов / А. А. Толкунов // Информационные системы и технологии. – 2014. – № 1(81). – С. 47–58.

3. Толкунов, А. А. Математические основы и программная архитектура системы оперативной аналитической обработки комментариев к обсуждаемым законопроектам / А. А. Толкунов // Системы управления и информационные технологии. – 2013. – № 3.1(53). – С. 173–177.

Статьи в научных журналах и сборниках трудов

4. Толкунов, А. А. О построении системы оперативной аналитической обработки комментариев в целях повышения эффективности процесса анализа результатов общественного обсуждения законопроектов / А. А. Толкунов // Информационные технологии моделирования и управления. – 2013. – № 4(82). – С. 339–351.

5. Толкунов, А. А. О погружении в массив комментариев с использованием типичных представителей мнений / А. А. Кукушкин, А. А. Толкунов // Труды Северо-Кавказского филиала Московского технического университета связи и информатики. – Ростов-на-Дону : ПЦ "Университет" СКФ МТУСИ, 2013. – 592 с. – С. 521–523.

(Соискателем описана проблема обработки массива текстовых комментариев в ходе общественного обсуждения законопроектов, предложен подход к снижению трудозатрат экспертов, приведены варианты методики определения типичного представителя класса массива комментариев, позволяющие минимизировать время обработки за счёт ознакомления с одним текстом из класса, содержащим основную мысль по корректировке текста статьи законопроекта).

6. Толкунов, А. А. Оперативная аналитическая обработка мнений в ходе общественного обсуждения законопроектов / А. А. Толкунов // Актуальные проблемы развития технологических систем государственной охраны, специальной связи и специального информационного обеспечения : VIII Всероссийская межведомственная научная конференция : материалы и доклады (Орёл, 13–14 февраля 2013 г.). В 10 ч. Ч. 8 / под общ. ред. В. В. Мизерова. – Орёл : Академия ФСО России, 2013. – 139 с. – С. 86–89.

7. **Толкунов, А. А.** Модель гиперкуба для представления текста в системе понимания и анализа текстовой информации / А. А. Кукушкин, А. А. Толкунов // Проблемы развития технологических систем государственной охраны, специальной связи и специального информационного обеспечения : Седьмая научно-практическая конференция : сборник материалов (Орёл, 3–4 марта 2011 г.). Ч. 4. – Орёл : Академия ФСО России, 2011. – 186 с. – С. 92–95.

(Соискателем исследована коммуникативная функция текста при передаче сообщения, предложены модель анализа текстов на основе гиперкуба и структура гиперкуба для представления смысла текста).

8. **Толкунов, А. А.** Применение гиперкуба как модели представления текста при решении информационно-аналитических задач / А. А. Кукушкин, А. А. Толкунов // Проблемы развития технологических систем государственной охраны, специальной связи и специального информационного обеспечения : Седьмая научно-практическая конференция : сборник материалов (Орёл, 3–4 марта 2011 г.). Ч. 4. – Орёл : Академия ФСО России, 2011. – 186 с. – С. 89–92.

(Соискателем исследованы способы представления содержания текстовых документов в виде гиперкуба – модели представления информации для оперативной аналитической обработки документов, проанализировано современное состояние разработок в указанной области и предложена модель обработки текстов на основе гиперкуба).

Свидетельства об официальной регистрации программ для ЭВМ

9. **Толкунов, А. А.** Классификатор функциональных стилей текстов "Style-Classifier" / А. А. Овсянников, А. А. Толкунов // Свидетельство об официальной регистрации программы для ЭВМ № 2006610680 от 20.02.2006 г.

(Соискателем осуществлены разработка и программная реализация алгоритма классификации функциональных стилей текстов, формирование эталонов классов, выбор метрики, обоснование оптимальной структуры признакового множества).

10. **Толкунов, А. А.** Программа атрибутивного анализа текстовых комментариев интернет-пользователей к проектам правовых актов "Comment Eliminator" / А. А. Кукушкин, А. А. Толкунов, Е. А. Шевченко // Свидетельство о государственной регистрации программы для ЭВМ № 2012616109 от 04.07.2012 г.

(Соискателем осуществлена разработка алгоритма атрибутивного анализа, структуры лингвистического обеспечения и формата выходных данных).

Подписано в печать 18.04.2014 г. Формат 60 × 84/16. Печать офсетная.
Усл. печ. л. 1,5. Тираж 100 экз. Заказ № 119
Отпечатано в типографии Академии ФСО России
302034, г. Орёл, ул. Приборостроительная, 35