УДК 004.912

ОНТОЛОГИЧЕСКИЕ РЕСУРСЫ И ИНФОРМИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА В ПРЕДМЕТНОЙ ОБЛАСТИ «БЕЗОПАСНОСТЬ»

Н.В. Лукашевич[1], Б.В. Добров[2], А.М. Павлов[3], С.В. Штернов[4]

Научно-исследовательский вычислительный центр МГУ имени М.В. Ломоносова, Москва, Россия
[1]louk_nat@mail.ru, [2]dobrov_bv@mail.ru, [3]pavlov.andrew.m@gmail.com, [4]shternov@gmail.com

Аннотация

В статье рассматривается подход к описанию широкой области национальной безопасности как тезауруса (лингвистической онтологии) для автоматической обработки документов. Созданный Тезаурус по безопасности имеет модель представления тезауруса РуТез и не имеет аналогов в мире в такой широкой предметной области. Тезаурус по безопасности используется в специализированной информационно-аналитической системе, в том числе для автоматической текстовой классификации документов в соответствии с несколькими рубрикаторами, включая рубрикатор угроз, рубрикатор ценностей, рубрикатор региональных проблем и др. Используемая информационно-поисковая система NearIdx обеспечивает стандартные функции информационного поиска, а также даёт возможность задания запросов и поиска информации с использованием специализированных ресурсов, включая онтологию и рубрикаторы. Аналитический компонент NearIdx предоставляет возможности фасетного анализа, спектрально-фасетного анализа временных рядов, построения когнитивных схем и аналитических справок, в которых могут использоваться созданные лексико-терминологические ресурсы.

Ключевые слова: *тезаурус, онтология, национальная безопасность, информационный поиск, рубрикатор, автоматическая классификация, текстовая аналитика*.

Цитирование: *Лукашевич, Н.В.* Онтологические ресурсы и информационно-аналитическая система в предметной области «Безопасность» / Н.В. Лукашевич, Б.В. Добров, А.М. Павлов, С.В. Штернов // Онтология проектирования. – 2018. – Т. 8, №1(27). - С.74-95. – DOI: 10.18287/2223-9537-2018-8-1-74-95.

Введение

Настоящее время характеризуется огромными объёмами электронной текстовой информации, пополняемой ежедневно, включая новости и газетные публикации, аналитические отчёты, научные статьи, а также сообщения в социальных сетях. Такого рода электронные текстовые ресурсы дают новые возможности для выявления происходящих процессов и возникающих трендов в разных сферах деятельности [1-4].

Одной из важных сфер применения мониторинга текстовой информации является сфера национальной безопасности, для которой важно отслеживание и предупреждение реальных и потенциальных угроз. Термин «национальная безопасность» определяется в Стратегии национальной безопасности Российской Федерации[1] как состояние защищённости личности, общества и государства от внутренних и внешних угроз, при котором обеспечиваются реализация конституционных прав и свобод граждан Российской Федерации, достойные качество и уровень их жизни, суверенитет, независимость, государственная и территориальная целостность, устойчивое социально-экономическое развитие Российской Федерации. Национальная безопасность включает в себя оборону страны и все виды безопасности, предусмотренные

[1]http://www.consultant.ru/document/cons_doc_LAW_191669/61a97f7ab0f2f3757fe034d11011c763bc2e593f/

Конституцией Российской Федерации и законодательством Российской Федерации, прежде всего государственную, общественную, информационную, экологическую, экономическую [5], транспортную, энергетическую безопасность, безопасность личности.

В данной статье рассматриваются возможности для определения угроз и трендов в разных сферах предметной области «Безопасность» на основе специализированных лексико-терминологических ресурсов: лингвистических онтологий и специализированных рубрикаторов. Применяемая лингвистическая онтология представляет собой расширенную версию тезауруса РуТез [6]. Созданные рубрикаторы содержат в себе категории существующих ценностей, т.е. того, что нужно защищать, и предполагаемых угроз, выявленных как по нормативным документам, так и на основе анализа текстов средств массовой информации.

Созданные лексико-терминологические ресурсы используются для автоматической классификации документов, а также в рамках специализированной информационно-аналитической системы. В статье рассматриваются инструменты, которые входят в информационно-аналитическую систему и дают дополнительные возможности для проведения мониторинга поступающей информации.

1 Методы автоматической обработки документов в сфере безопасности

В литературе представлено несколько направлений обработки текстов на естественном языке в сфере национальной и международной безопасности.

Многие работы посвящены анализу экстремистских сообщений в социальных сетях. Часть исследований проявлений терроризма и экстремизма в Twitter посвящена деятельности ИГИЛ (Исламское государство Ирака и Леванта). Известно, что эта террористическая организация активно работает с множеством социальных сетей [7-8]. В частности, ИГИЛ и связанные с ним организации поддерживают большое количество учётных записей Twitter на многих языках для распространения своих идей. В работе [9] дан анализ сообщений в социальных сетях людей, которые присоединились или попытались присоединиться к ИГИЛ. В большинстве случаев они выражали сильные антиамериканские и антизападные настроения задолго для присоединения к этой организации.

В статье [10] описывается корпус, содержащий 100 текстов, написанных исламистами. В частности, из собраний хадисов (исламских религиозных текстов) были извлечены фрагменты о войне, неверующих и наказаниях (64 текста). Кроме того, корпус содержит сообщения из исламских блогов, а также статьи из исламистского журнала Inspire. Эти тексты были размечены на нескольких уровнях, включая синтаксические, временные и другие аннотации.

Поскольку одним из предпосылок поддержки терроризма является резко негативное отношение к определённым явлениям или группам людей, то так называемые заявления «ненависти» (“hate”) требуют особого внимания. В работе [11] сообщения в блогах классифицируются не только на классы выраженных эмоций (положительные, отрицательные, нейтральные), но и в зависимости от типа обсуждаемого действия (отрицательный - *кровопролитие, жестокость*, или положительный – *помощь, поддержка*). Квок и Ван [12] описывают собранный сбалансированный набор из 24,5 тыс. твитов, которые классифицируются как расистские и нормальные. В [13] представлен корпус из 16 000 твитов, в которых 3,3 тыс. твитов обозначены как сексистские, а 1,9 тыс. – расистские.

В работе [14] указывается, что автоматическое обнаружение заявлений, которые разжигают ненависть по отношению к некоторым группам населения (что является одним из признаков экстремизма), осложняется следующими факторами:

- сообщение не всегда может быть найдено на основе простого набора ключевых слов, потому что некоторые слова намеренно искажены (чтобы избежать обнаружения);

- любые фиксированные списки оскорбительных слов постоянно требуют добавления;
- заявления о ненависти могут быть написаны на прекрасном литературном языке;
- преступление и ненависть могут пересекать границы предложения, когда объект, на который направлено высказывание, находится в другом предложении;
- использование сарказма.

Шмидт и Виганд [15] предоставляют обзор существующих подходов к обнаружению сообщений о ненависти. Отмечается, что такие подходы основаны на применении методов машинного обучения на основе нескольких групп признаков.

В сфере автоматической обработки текстов по информационной безопасности Лим и др. [16] обсуждают создание базы данных для аннотирования текстов описаний вредоносных программ. Структура аннотаций вводится на основе словаря MAEC для определения характеристик вредоносного программного обеспечения [17], а также базы данных. Авторы планируют использовать базу данных для создания моделей, которые потенциально могут помочь исследователей кибербезопасности в своих усилиях по сбору и анализу данных. Горохов и др. [18] изучают применение свёрточной нейронной сети для обнаружения аномалий в данных электронной почты.

Несколько проектов посвящены глобальному мониторингу происходящих в мире событий, что необходимо для понимания существующих проблем и формирования подходов к их разрешению, например, Международная система раннего предупреждения о кризисах (ICEWS), поддерживаемая Lockheed Martin и Global Data on Events Language and Tone (GDELT) [19-20]. Для прогнозирования эскалации конфликтов [19] использовались ручные и автоматически собранные данные о событиях. В ICEWS используются также статистические и агентные модели и утверждается, что точность прогнозирования составляет 80%. Система GDELT была использована для отслеживания сообщений, вызывающих ненависть, после голосования по поводу отделения Великобритании от Евросоюза [20].

2 Лингвистическая онтология в области безопасности

Для представления знаний в сфере безопасности используется модель комплекса лингвистических онтологий РуТез, которые представляют собой формализованные информационно-поисковые тезаурусы, ориентированные на автоматическую обработку больших текстовых коллекций и поддержку решения задач текстовой аналитики [6, 21]. В модели учитываются три парадигмы описания знаний в широких предметных областях: информационно-поисковые тезаурусы, тезаурусы типа WordNet, формальные онтологии.

В настоящее время существует несколько крупных русскоязычных лингвистических онтологий (тезаурусов).

- Тезаурус РуТез, содержащий слова и фразы литературного русского языка вместе с терминами так называемой общественно-политической области [22].
- Тезаурус РуТез-lite, опубликованная версия РуТез[1], который исследователи могут получить бесплатно для некоммерческого использования [23].
- Общественно-политический тезаурус, содержащий слова и выражения литературного языка и термины из общественно-политической области. Общественно-политическая область – это широкая область современных общественных отношений, представляющая повседневную жизнь современного общества и объединяющая многие профессиональные сферы, такие как политика, право, экономика, международные отношения, финансы, военные дела, искусство и другие. Особенностью этой предметной области является то, что

[1] http://www.labinform.ru/pub/ruthes/index.htm

термины этой предметной области обычно известны не только профессионалам, но и обычными людям [24], поскольку это сфера пересечения профессиональных и общих знаний о мире. Общественно-политический тезаурус может использоваться отдельно для обработки документов. В то же время он входит в состав трех крупных тезаурусов: РуТез, онтологии ОЕНТ и Тезауруса по безопасности.

- Онтология по естественным наукам и технологиям (ОЕНТ), которая включает в себя термины математики, физики, химии, геологии, астрономии и т. д., термины инженерных предметных областей (нефть и газ, электростанции, космические технологии, летательные аппараты и т. д.). ОЕНТ также включает Общественно-политический тезаурус, поскольку научно-технические проблемы могут обсуждаться вместе с политическими, экономическими, промышленными и другими вопросами [25].
- Тезаурус безопасности, который является расширением тезауруса РуТез и включает терминологию, связанную с социальными, национальными и религиозными конфликтами, экстремизмом и терроризмом, информационной безопасностью.

В таблице 1 приведены количественные характеристики вышеперечисленных ресурсов.

Таблица 1 - Лингвистико-терминологические ресурсы в формате тезауруса РуТез

Тезаурус	Число понятий	Число текстовых входов	Число отношений
РуТез	55275	170130	226743
РуТез-lite	31540	111559	128866
Общественно-политический тезаурус	41426	121292	161523
Онтология ОЕНТ	94103	262955	376223
Тезаурус по безопасности	66810	236321	271297

Тезаурус по безопасности представляет собой лингвистическую онтологию, т.е. онтологию, понятия которой опираются на значения существующих в языке слов и выражений. Каждое понятие онтологии имеет уникальное и однозначное имя. Каждое понятие онтологии связано с набором слов и выражений, посредством которых данное понятие может выражаться в тексте – текстовые входы понятия. Набор текстовых входов понятия может включать собственно синонимы, слова разных частей речи (так называемые дериваты), устойчивые словосочетания и другие типы выражений.

На рисунках 1-3 показан интерфейс разработки тезауруса. Верхняя левая форма содержит список понятий, нижняя левая форма показывает текстовые входы для выделенного понятия. Правая верхняя форма представляет отношения выделенного понятия, а нижняя правая форма показывает текстовые входы для соответствующих понятий. Можно видеть, что понятия снабжены многочисленными текстовыми вариантами, извлеченными из реальных текстов, например, понятие *импортная зависимость* (рисунок 1) может быть выражено в тексте как *зависимость от импорта* или *зависимость от импортных товаров*. На рисунке 3 показаны варианты языковых выражений, используемых для выражения понятий *атака отказа обслуживания* и *распределенная DOS-атака* в сфере компьютерной безопасности.

Как и в РуТез, в Тезаурусе по безопасности имеется четыре основных типа отношений [24]. Первый тип отношений – родовидовое отношение *ниже-выше*, представляет собой отношение класс-подкласс, обладает свойствами транзитивности и наследования.

Второй тип отношений – отношение *часть-целое*. Отношение используется не только для описания физических частей, но и для других внутренних сущностей понятия, таких как свойства или роли для ситуаций. Важным условием при установлении этого отношения является то, что понятия-части должны быть жёстко связаны со своим целым, то есть каждый пример понятия-части должен в течение всего времени своего существования являться ча-

стью для понятия-целого и не относиться к чему-либо другому. В этих условиях удаётся выполнить свойство транзитивности введённого таким образом отношения часть-целое, что очень важно для автоматического вывода в процессе автоматической обработки текстов. Кроме того, в таких условиях при появлении в тексте упоминания части можно автоматически выводить, что текст имеет отношение и к целому.

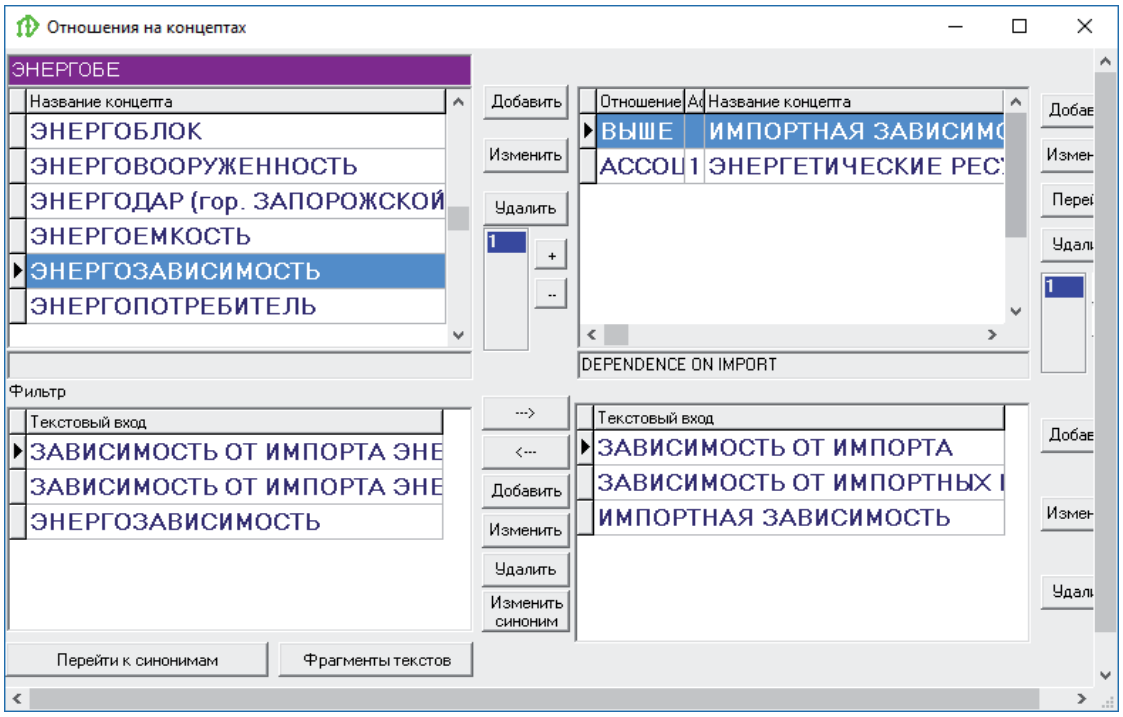


Рисунок 1 - Текстовые входы и отношения понятия *энергозависимость* в сфере экономической безопасности

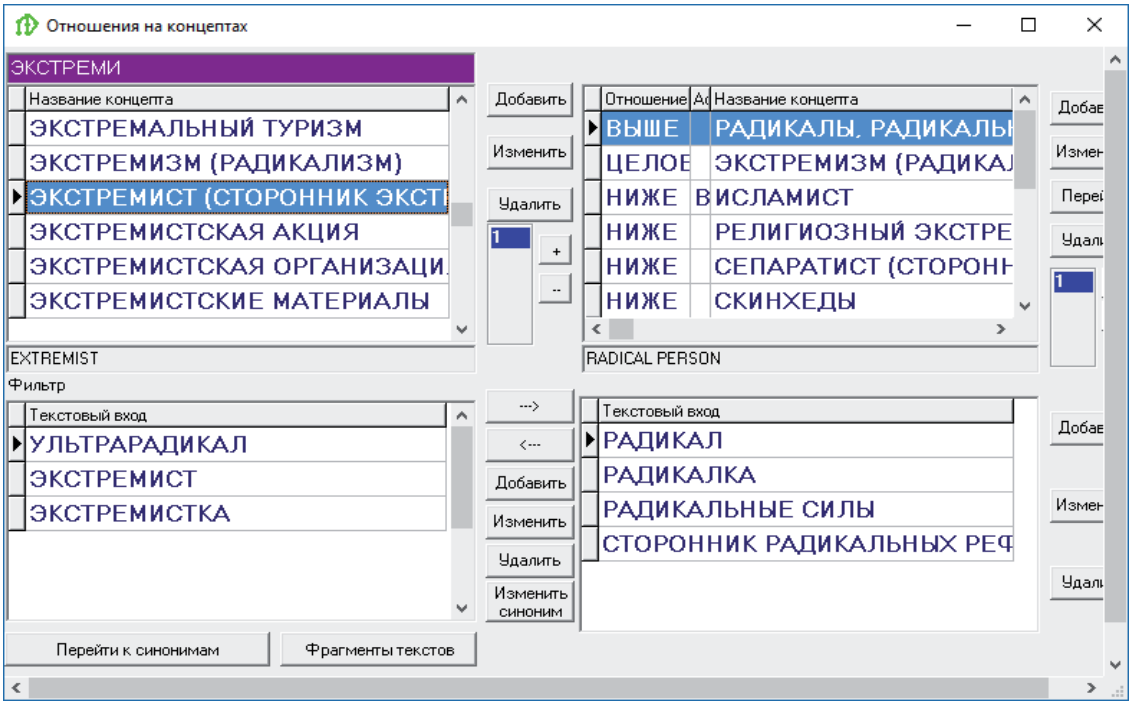


Рисунок 2 - Текстовые входы и отношения для понятия *экстремист* в сфере борьбы с экстремизмом

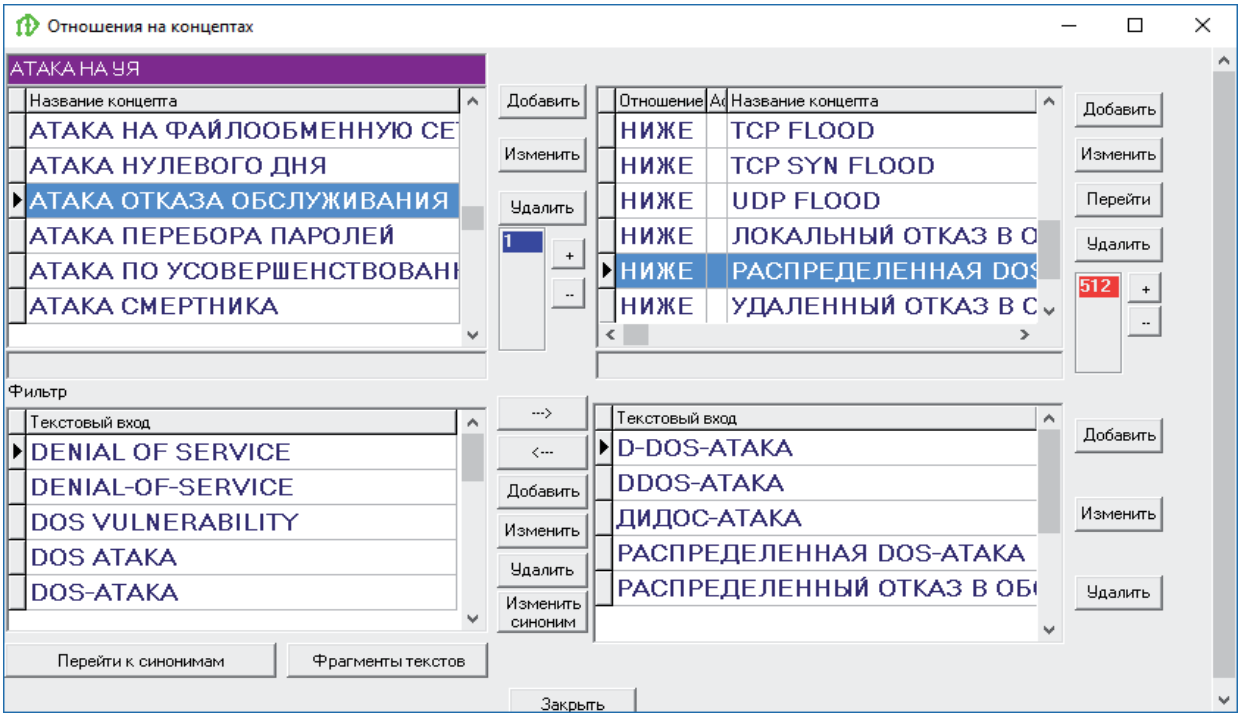


Рисунок 3 - Текстовые входы и отношения для понятия *атака отказа обслуживания* в сфере компьютерной безопасности

Ещё один тип отношения, называемого несимметричной ассоциацией $асц_2$ - $асц_1$, связывает два понятия, которые не находятся в отношениях часть-целое или класс-подкласс, и существование одного из понятий требует существования другого понятия. На рисунке 1 показан пример такой зависимости понятия *энергетическая зависимость* от понятия *энергоресурсы*. Также понятие *замещение импорта* появляется, когда существует понятие *импортной зависимости*. В сфере компьютерной безопасности можно привести пример понятия *атака на уязвимость,* существование которого зависит от понятия *уязвимость в программе*.

Последний тип отношений - симметричная ассоциация *асц* связывает, например, понятия, очень близкие по смыслу, но которые разработчики не решились соединить в одно понятие (предсинонимия).

Таким образом, система отношений в тезаурусах типа РуТез описывает наиболее существенные отношения понятий. Рисунок 4 показывает фрагмент иерархии понятий, относящихся к понятию *опасность, угроза*.

Тезаурус по безопасности как отдельный ресурс создаётся в течение последних пяти лет. Он пополняется на основе нескольких источников.

Во-первых, анализируется существующая нормативно-справочная литература, содержащая термины, относящиеся к сфере безопасности. Такие термины вносятся в тезаурус.

Во-вторых, создаются специализированные текстовые коллекции по конкретным подобластям сферы безопасности, из которых производится автоматическое извлечение кандидатов в термины на основе синтаксических, лексических и синтаксических признаков [5]. Извлечённые слова и выражения просматриваются терминологами и вносятся в тезаурус в виде новых понятий или как текстовые входы к существующим понятиям.

В-третьих, поступающие новости из средств массовой информации по тематике безопасности обрабатываются на основе текущей версии тезауруса (см. раздел 3 и рисунок 5), при этом может быть выявлена нехватка терминов, которые добавляются в тезаурус.

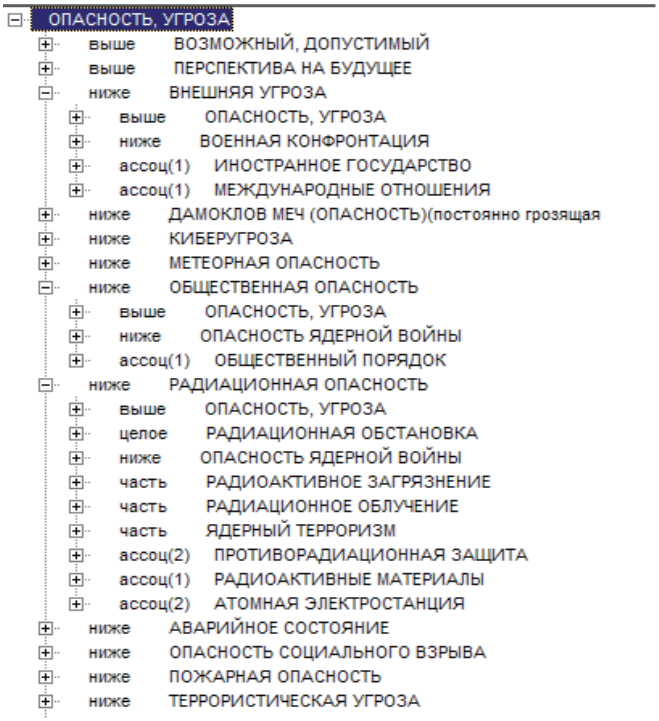


Рисунок 4 - Фрагмент иерархии понятий, относящихся к понятию *опасность, угроза*

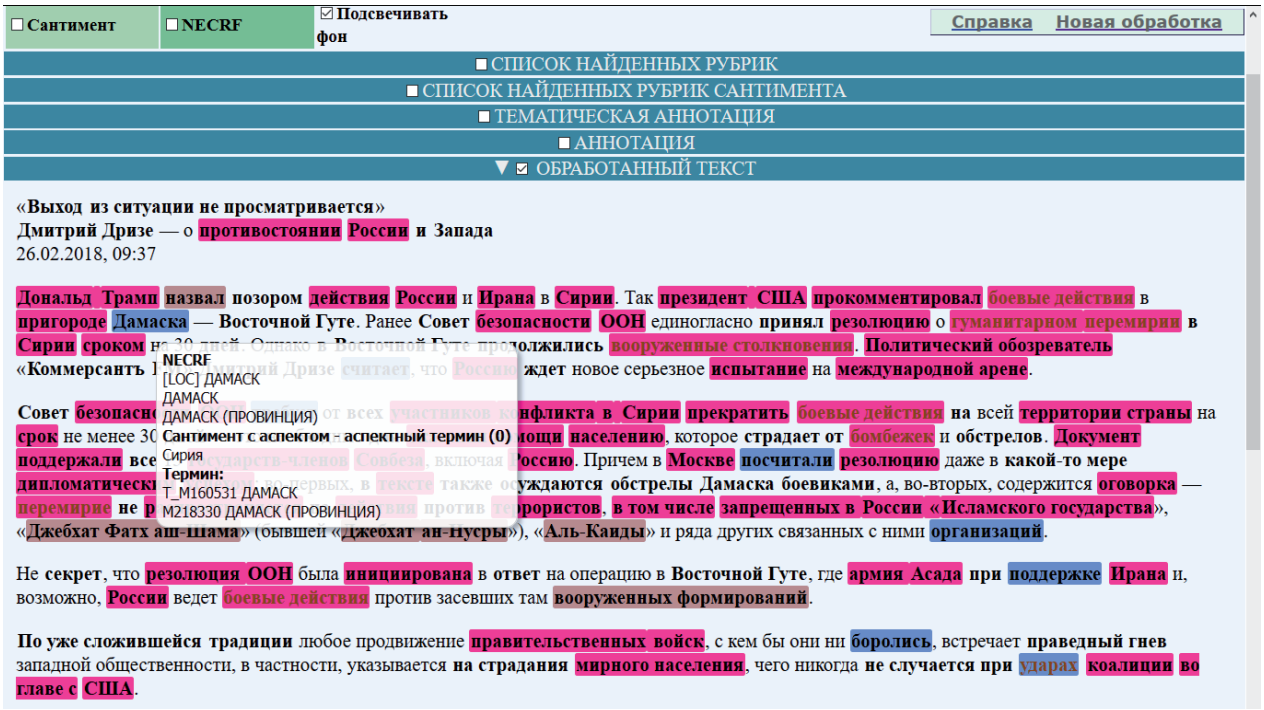


Рисунок 5 - Пример сопоставления текста с тезаурусом

3 Автоматическая обработка текстов на основе тезауруса

Обработка текстов на основе тезауруса включают в себя следующие этапы.

- Графематический и морфологический анализ, который переводит разные формы слова к единой словарной форме (лемме).

- Терминологический анализ документа, состоящий в автоматическом сопоставлении текста с тезаурусом на основе последовательностей лемм. На рисунке 5 показан результат терминологического анализа текста «О противостоянии России и Запада»[1]. Выделенные слова и словосочетания найдены в Тезаурусе по безопасности.
- Автоматическое разрешение многозначности слов. Фрагменты текста, выделенные на рисунке 5 коричневым и синим цветом, означают обнаруженные многозначные термины, т.е. термины, которые соответствуют разным понятиям тезауруса. Так, имя *Дамаск* может относиться к столице Сирии или сирийской провинции. Разрешение многозначности слов производится на основе сопоставления контекста неоднозначного слова в документе и списка близких по тезаурусу слов и словосочетаний [6]. Так, в тексте примера разметка ”T_M” означает, что выбрано значение *Дамаска* как столицы.
- Автоматическое построение тематического представления текста заключается в том, что близкие по смыслу понятия группируются в так называемые тематические узлы, которые затем распределяются по их значимости для текста: основные и локальные. Значимость тематического узла учитывается в формируемом весе понятия для формирования концептуального индекса документа. На рисунке 6 показаны извлечённые из вышеупомянутого текста тематические узлы, где каждый следующий узел показан с отступом по отношению к предыдущему. Видно, что в один из крупных узлов собрались различные понятия, связанные с демографической обстановкой.

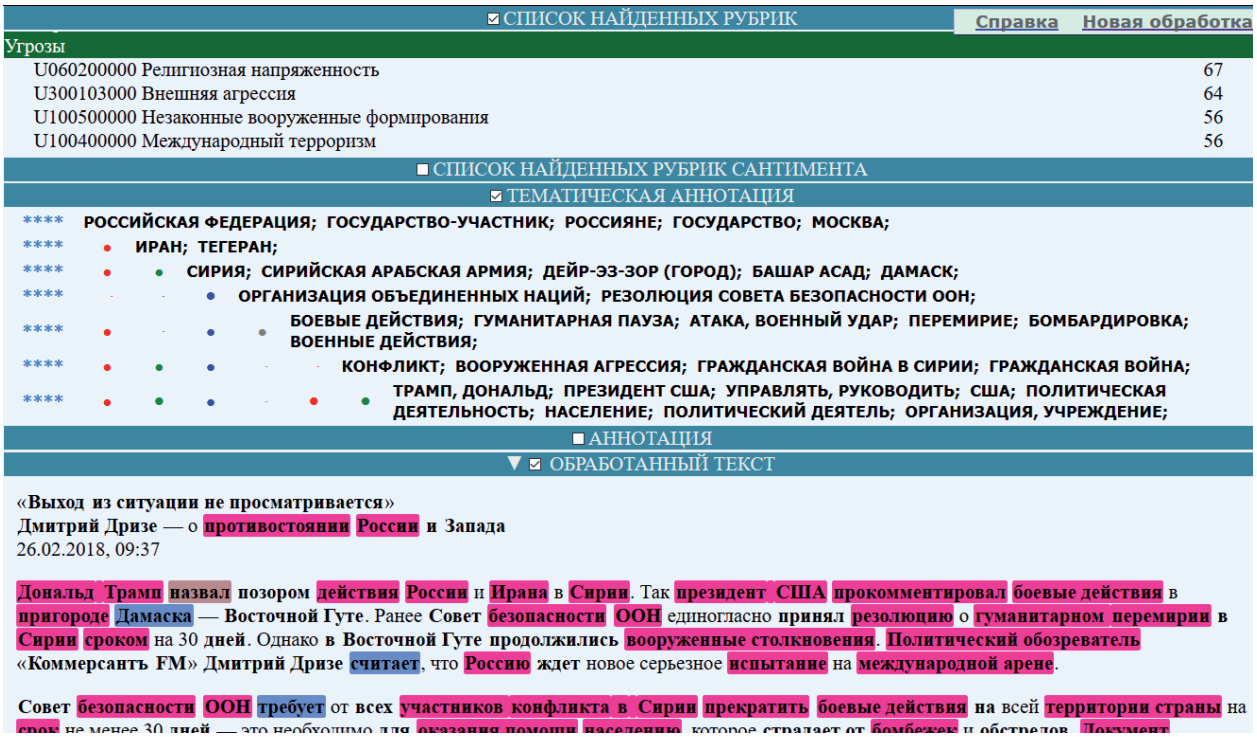


Рисунок 6 - Пример автоматически проставленных рубрик и тематических узлов

- Формирование концептуального индекса документа. Концептуальный индекс документа состоит из понятий, найденных в документе, и их весов. Вес понятия учитывает значимость соответствующего узла темы и частоту понятия в документе. В тексте примера понятие *гуманитарная пауза* было явно упомянуто только один раз в тексте, и соответствующее понятие может получить слишком низкий вес, если учитывать только частоту

[1] https://www.kommersant.ru/doc/3558551

упоминания, но при поддержке основного тематического узла «боевые действия» вес понятия *гуманитарная пауза* становится значительно выше.

- Автоматическая рубрикация по заданному рубрикатору. На рисунке 6 показаны рубрики, которые были автоматически получены для текста примера: *Религиозная напряженность, Внешняя агрессия, Международный терроризм* и др.
- Автоматическое аннотирование и др.
- Результаты обработки документа, включая пословный индекс, концептуальный индекс, вычисленные рубрики и т.п. загружаются в информационно-аналитическую систему.

Результаты обработки отдельного текста используются для построения компонентов различных информационных систем – информационного поиска, агрегирования новостных потоков (кластеризация), составления аналитических отчётов (различные типы обзорных рефератов) и т.д.

4 Автоматическая рубрикация текстов на основе тематического представления

Основной современной технологией автоматической классификации текстов является подход на основе машинного обучения. Этот подход предполагает наличие текстовой коллекции достаточного объёма для обучения алгоритмов. Однако в новой сложной задаче рубрикации текстов даже собственно рубрикатор, т.е. система категорий для рубрикации, может отсутствовать и должен быть создан с нуля или с использованием существующих похожих рубрикаторов.

В таких условиях более приемлемыми являются методы автоматической рубрикации, основанные на знаниях, т.е. на ручных правилах присвоения рубрик. Создавая рубрикатор и правила вывода рубрик в широкой предметной области, необходимо использовать поддержку тезауруса в написании правил, потому что тезаурус позволяет работать не с отдельными словами и выражениями, а с понятиями и подструктурами тезауруса [24].

Используемая процедура автоматической рубрикации текстов базируется на автоматически построенном тематическом представлении документов, которое моделирует основную тему и подтемы документа наборами (тематическими узлами) близких по смыслу понятий, упомянутых в документе [6, 24]. Такая основа рубрикации даёт возможность обрабатывать тексты разных типов и размеров: нормативные акты, газетные статьи, новостные сообщения, научные публикации в области гуманитарных наук, социологические опросы.

При создании лингвистического профиля рубрикатора каждая рубрика R описывается (см. рисунок 7) дизъюнкцией альтернатив, каждый дизъюнкт представляет собой конъюнкцию $R = \bigcup_i D_i \;;\; D_i = \bigcap_j K_{ij}$. Конъюнкты в свою очередь описываются экспертами с помощью так называемых «опорных» понятий. Для каждого опорного понятия задаётся правило его расширения *f(·)*, определяющее, каким образом вместе с опорным понятием учитывать подчинённые ему по иерархии понятия: без расширения (обозначается символом «N»), полное расширение по дереву иерархии (символ «E»), расширение только по родовидовым связям (символ «L»), расширение по всем видам отношений на один уровень иерархии (символ «W»), расширение на один уровень иерархии, не включая отношения *ниже* (символ «V»).

Например, один из компонентов описания рубрики «Религиозная напряжённость» представляет собою конъюнкцию, первый элемент которой понятие *конфликт* без расширения (N), а второй элемент - понятия *религия* и *верующий* с расширением по видам (L), т.е. данная рубрика будет выводиться, если в тексте встречается упоминание понятия *конфликт* с любым понятием, относящимся к религиям, включая различные конфессии, верующих, религи-

озные догмы, ритуалы и т.п. Таким образом, с помощью тезауруса преодолевается проблема трудоёмкости инженерных методов рубрикации: если тезаурус уже существует, то описание рубрик, вывод которых основан на большом количестве разных слов, делается достаточно быстро за счёт иерархической организации тезауруса. Описание эксперимента по измерению времени создания описаний для рубрик рубрикатора и оценки качества получившейся системы рубрикации приводится в [24].

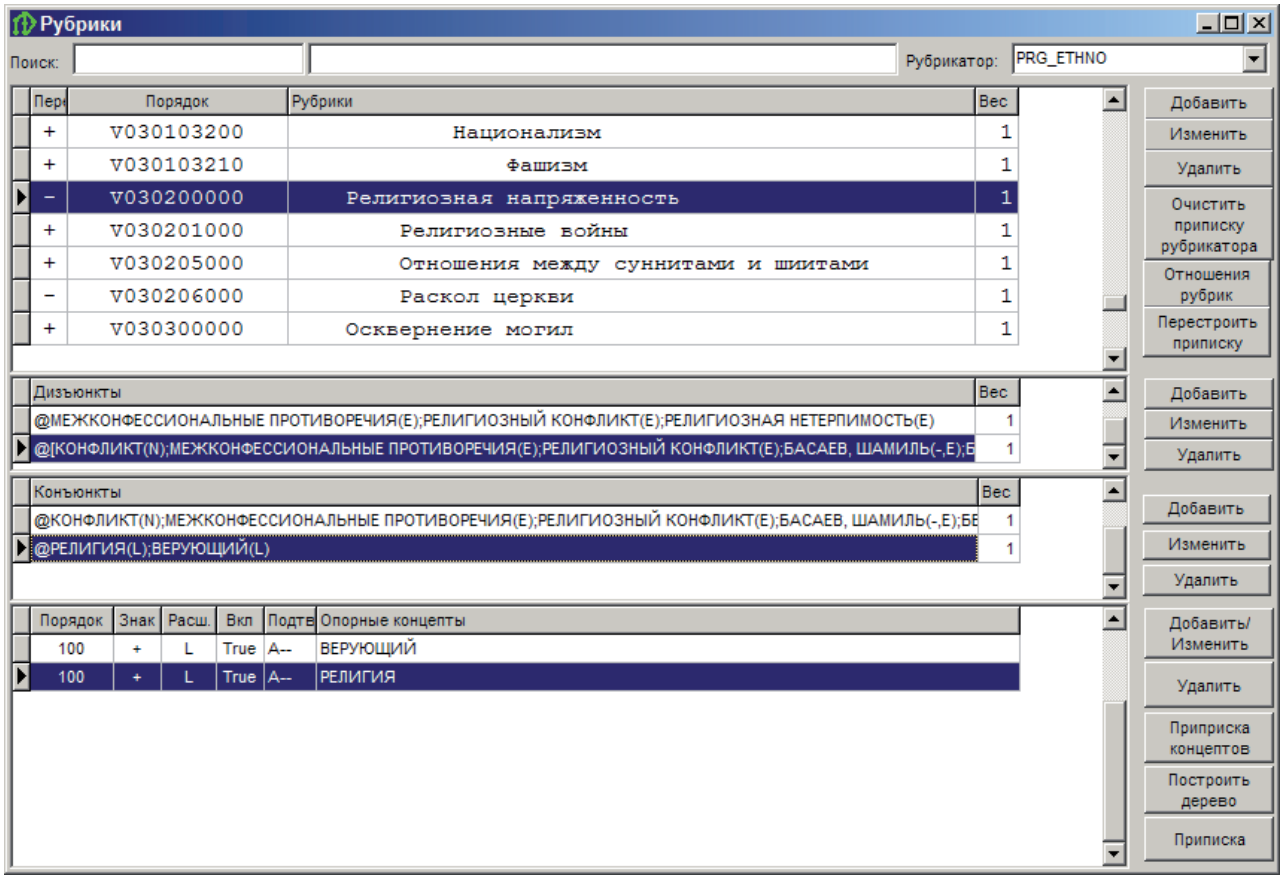


Рисунок 7 - Интерфейс описания содержания рубрик с помощью понятий тезауруса. Представление рубрики «Религиозная напряженность»

Вес конъюнкта зависит от максимального веса входящего в него понятия онтологии. Вес дизъюнкта предназначен учитывать не только сумму весов составляющих его конъюнктов, но и меру близости конъюнктов в тексте. Вес рубрики представляет собой максимум весов входящих в описание рубрики альтернатив. В случае имеющихся иерархических связей между рубриками оценка релевантности нижестоящих рубрик переносится на вышестоящие, так что при запросе по вышестоящей рубрике будут выходить и документы, к которым были приписаны нижестоящие рубрики.

Алгоритм рубрицирования работает следующим образом. Для всех понятий тезауруса, найденных в тексте, определяется множество рубрик, которые могут быть определены в тексте. В результирующем множестве остаются рубрики, вес которых превосходит задаваемый заранее для коллекции порог.

Для предметной области «Безопасность» разработано несколько рубрикаторов:

- рубрикатор угроз, описывающий существующие угрозы национальной безопасности (188 категорий, 5-уровневая иерархия);

- рубрикатор ценностей, представляющий индивидуальные, социальные и государственные ценности, например, свобода, демократия, права человека, семейные ценности и т. д. (109 категорий, иерархия 4 уровня);
- рубрикатор этно-конфессиональных отношений (94 категории, иерархия 4 уровня);
- рубрикатор региональных проблем (84 категории, 2-уровневая иерархия);
- рубрикатор регионов (325 регионов: субъекты Российской Федерации и иностранные государства).

5 Информационно-аналитическая система

Основным инструментом анализа больших коллекций текстовых документов является информационно-аналитическая система. В результате обработки отдельного текста выделяется большое количество объектов разных типов, которые традиционно привязываются к словопозициям (порядковому номеру слова в тексте). Результаты обработки загружаются в поисковые индексы информационно-поисковой системы.

Использование онтологий позволяет организовать работу аналитика более эффективно, как в части задания информационных потребностей, так и в части интерпретации получаемых от поисковой машины результатов.

5.1 Базовый поиск и организация поисковых индексов

В качестве поисковой машины авторы используют noSQL базу данных NearIdx 9.0, развиваемую Лабораторией информационных исследований совместно с Научно-исследовательским вычислительным центром МГУ. Традиционно используются два вида поисковых индексов: обратный и прямой, которые хранятся в специальных noSQL базах данных, поскольку накладные расходы на перестроение поисковых индексов для текстовых коллекций делают использование реляционных баз данных неэффективным.

Обратный (инверсный) индекс хранит информацию, организованную от элемента словаря данных к документу. Для каждого элемента данных E_i поддерживаются структуры данных по документам d_j, оценка релевантности документа rank (E_i, d_j) и вектор словопозиций [$positions_{ijk}$].

$$[\ E_i\ \ [d_j,\ rank(E_i,\ d_j),\ [\ positions_{ijk}\ ]\ \ ]\ \ ]$$

Обратный индекс используется для поиска документов. В запросе пользователя к системе определяются элементы словарей данных. Для каждого элемента с жёсткого диска в оперативную память извлекаются указанные структуры, которые пересекаются между собой, образуя список документов, куда входят либо все элементы запроса при «строгом» поиске, либо не все при задании условий «нежёсткого/нечёткого» поиска.

Прямой индекс хранит ту же информацию, но организованную по документам:

$$[\ d_j\ \ \ [E_i\ ,\ \mathrm{rank}(E_i,\ d_j),\ [\ positions_{ijk}\ ]\ \ ]\ \ ]$$

Прямой индекс используется, например, для подсветки релевантных запросу фрагментов документов, что позволяет ускорить процесс отбора нужных пользователю документов. Также прямой индекс является основным средством поддержки для продвинутых средств анализа поисковой выборки, когда для первоначального запроса получена группа документов и требуется выбрать наиболее значимые элементы в данной выборке.

Для ускорения поддержки анализа временных рядов используются вспомогательные индексы по датировке публикации документов:

$$[\ E_i\ \ [datetime_j\ ,\ \mathrm{doc\_count}(E_i\ ,\ \mathrm{rank}(E_i,\ d_j)>r_0)\ ]\ \ ]$$

Особенностью NearIdx является тесная интеграция с результатами обработки автоматической лингвистической обработки текстов (АЛОТ). Поддерживаются специальные поисковые индексы по словопозициям вхождения текстовых входов понятий лингвистических онтологий, в том числе с учётом иерархии понятий, а также пословные индексы, ассоциированные с описанием используемых рубрикаторов. Также поддерживаются индексы по выделенным именованным сущностям, тонально окрашенным словам и выражениям, прямой и косвенной речи и т.п.

Отметим, что поисковый индекс по понятиям с иерархией всего примерно в четыре раза превосходит по объёму индекс по понятиям без учёта иерархии (поиск по синонимам) и примерно равен по размеру пословному индексу с учётом морфологии.

Условия поиска задаются либо словами на обычном (естественном) языке, либо в следующем формате: /Словарь_данных= «Элемент данных». Например, запись запроса "Наркотики в Российской Федерации" по данным публикаций октября 2017 года с использованием понятий онтологии будет выглядеть следующим образом (регистр не важен):

/Термин_расш=«НАРКОТИК»

/Термин_расш=«РОССИЙСКАЯ ФЕДЕРАЦИЯ»

/Дата_док=«01.01.2017-31.10.2017»

Данный запрос позволит найти документы с упоминанием всех видов наркотиков и относящихся к наркотикам понятий, указанных в тезаурусе, с учётом всех регионов Российской Федерации.

При написании запроса допустимы логические операторы (отсутствие оператора эквивалентно логическому «И»), скобки для задания очерёдности обработки условий запроса, а также дополнительные контекстные операторы (поиск внутри одного или нескольких предложений, с начала или в конце текста, в окне заданного размера и т.д.).

Ключевым вопросом при обработке больших текстовых коллекций является время обработки запросов, которое при указанной организации определяется временем поиска и считывания информации с жёсткого диска. Время поиска по простым запросам по обратному индексу для коллекции 10 миллионов документов на стандартных современных компьютерах составляет 1-2 секунды.

Для поисковой машины недостаточно просто определить список релевантных документов. Необходимо сформировать выдачу, желательно с подсветкой релевантных фрагментов в найденных документах. Для сложно организованных поисковых индексов, например, для подсветки по рубрикам, требуется обращение к прямому индексу.

При этом интегральная (аналитическая) обработка первых 200 документов поисковой выдачи по прямому индексу составляет 0.3 секунды, 5 секунд для первых 2000 документов, а для анализа больших поисковых выборок может потребоваться значительное время. Поэтому стандартным решением для современных поисковых машин, в том числе для NearIdx, является очень быстрое получение именно первой страницы поисковой выдачи, но не всех страниц выдачи. Часто получение следующей страницы поисковой выдачи требует исполнения отдельного запроса.

5.2 Фасетный анализ

Для выборки документов доступны средства фасетного анализа по словарям данных. Можно установить количество наиболее релевантных запросу документов, по которым будет рассчитываться «информер» – специальная информационная панель, содержащая наиболее частотные или характерные для документов выборки элементы словаря данных. При этом возможны информеры по онтологии (см. рисунок 8) или по различным классификаторам (см. рисунок 9).

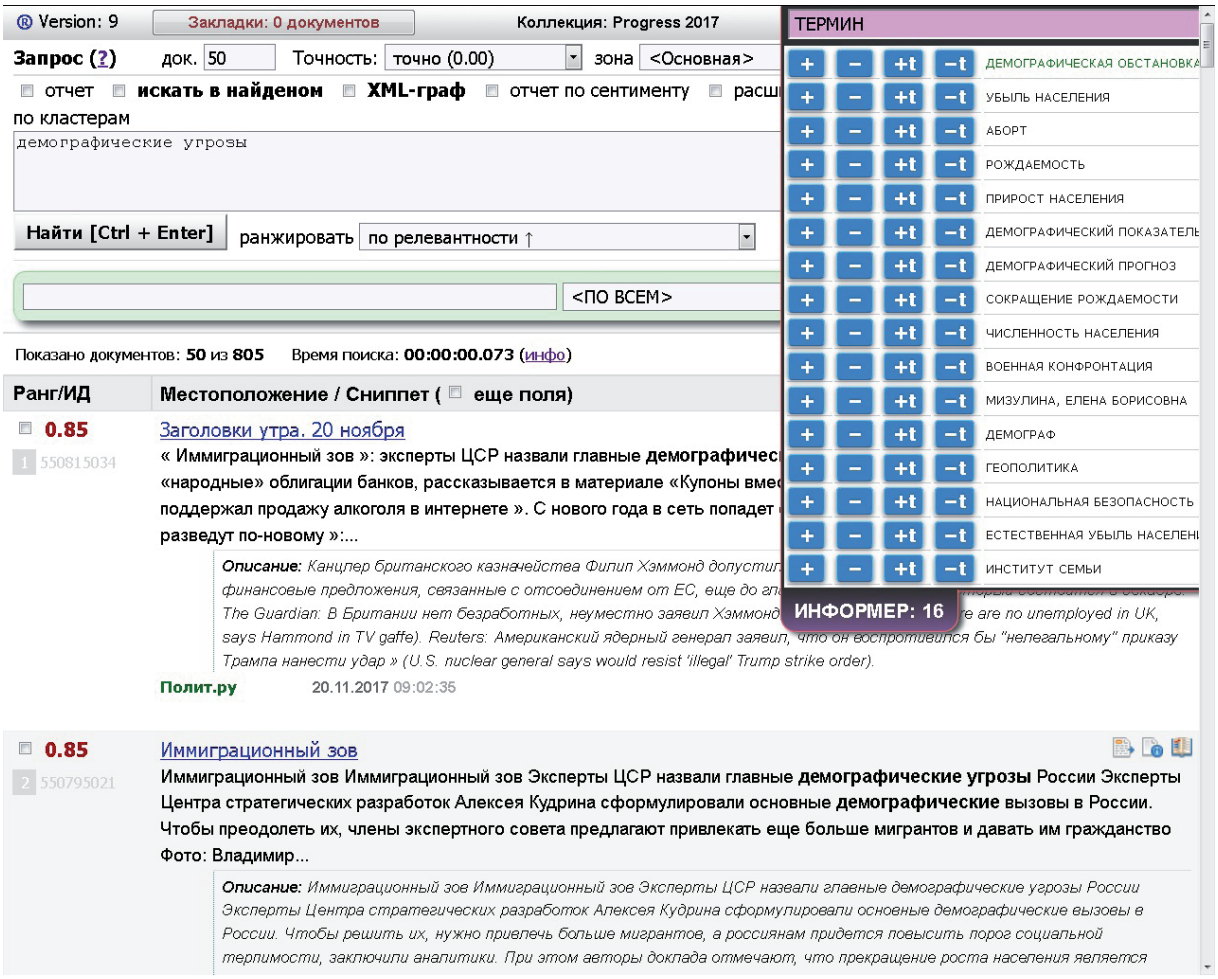


Рисунок 8 - Информер показывает наиболее характерные понятия онтологии для текущего запроса «Демографические угрозы»

а) Информер по упоминаемым регионам

б) Информер по упоминаемым угрозам

в) Информер по упоминаемым ценностям

Рисунок 9 - Информеры по разным классификаторам для запроса «Демографические угрозы»

На рисунке 9 показаны наиболее характерные категории классификаторов для запроса «Демографические угрозы»: регионы (а), угрозы (б), ценности (в). Кнопки «+» и «-» позволяют «одним кликом мыши» добавить соответствующее условие в запрос.

Обеспечение полноты поиска является фундаментальной проблемой из-за естественного разнообразия языка. Поиск по понятиям онтологии равносилен поиску с учётом синонимии, поиск по понятиям онтологии с учётом расширения по иерархии (когда релевантными считаются документы, содержащие синонимы выбранного понятия или его подчинённых по иерархическим связям) может быть весьма эффективным: запрос выглядит компактно, а результаты полны. Однако опыт эксплуатации различных информационно-поисковых систем выявил проблемы пользователей, которым трудно подобрать нужное понятие в большой понятийно-терминологической сети.

Реализация информера по понятиям онтологии (информер «Термины») позволяет эффективно решать данную проблему в динамике общения с информационно-поисковой системой. Действительно, пользователь может быстро исполнить запрос на естественном языке по интересующей его теме. Результаты поиска и информеры получаются очень быстро. При этом нужные понятия содержатся в информере. Далее достаточно одного-двух кликов «мыши» для выбора нужных понятий, которые и обеспечивают полноту поиска.

5.3 Спектрально-фасетный анализ временных рядов

Спектрально-фасетный анализ временных рядов темпорального распределения документов поисковой выдачи реализуется программным обеспечением SpectralView. Результаты выводятся в виде графика количества публикаций по разным запросам в зависимости от времени, при необходимости с требуемым уровнем агрегации (по часам/ дням/ неделям/ месяцам).

Автоматически, аналогично XYZ-статистикам, рассчитываются наиболее значимые факторы, графики которых наносятся на основной график. Для потоков текстовых документов задача определения XYZ-статистики естественным образом аналогична задаче выявления основных факторов, которые действуют на всём времени выборки документов или, наоборот, объясняют появление тех или иных пиков на временном графике.

Имеется возможность наложить несколько иерархий лингвистической онтологии, чтобы осуществлять выбор статистик только из заданных иерархий. Этим удобно пользоваться, например, чтобы выделять наиболее значимые именованные сущности: регионы, города и т.д. Для аналитика также полезна фильтрация факторов по иерархиям абстрактных понятий, например, по видам преступлений, видам болезней, видам зданий и т.п. Для удобства некоторые линии графиков можно выделять, другие, наоборот, затенять.

На рисунке 10 приведены данные анализа выборки (по корпусу российских новостей за 2017 год, 9 миллионов документов) по запросу: /Рубрика=«ЭКОЛОГИЧЕСКИЕ УГРОЗЫ». Как нетрудно видеть, резонанс обсуждения по теме в значительной мере формируется текстами, в которых обсуждаются пожары и лесные пожары.

Для любой точки графика можно, быстро исполнив дополнительный запрос, получить список документов, релевантный исходному запросу, выбранному элементу словаря данных и опубликованному в выбранном временном интервале. В частности, легко узнать, что пик на рисунке 10 в конце апреля объясняется публикациями по годовщине аварии на Чернобыльской АЭС 26.04.1986 г.

На рисунке 11 нанесены графики распределения по субъектам РФ публикаций с упоминанием тематики лесных пожаров (запрос /Термин_расш=«ЛЕСНОЙ ПОЖАР»). Согласно этим данным наибольший резонанс в марте-августе 2017 года вызывали лесные пожары в восточной Сибири – Иркутской области, Бурятии, Забайкальском крае.

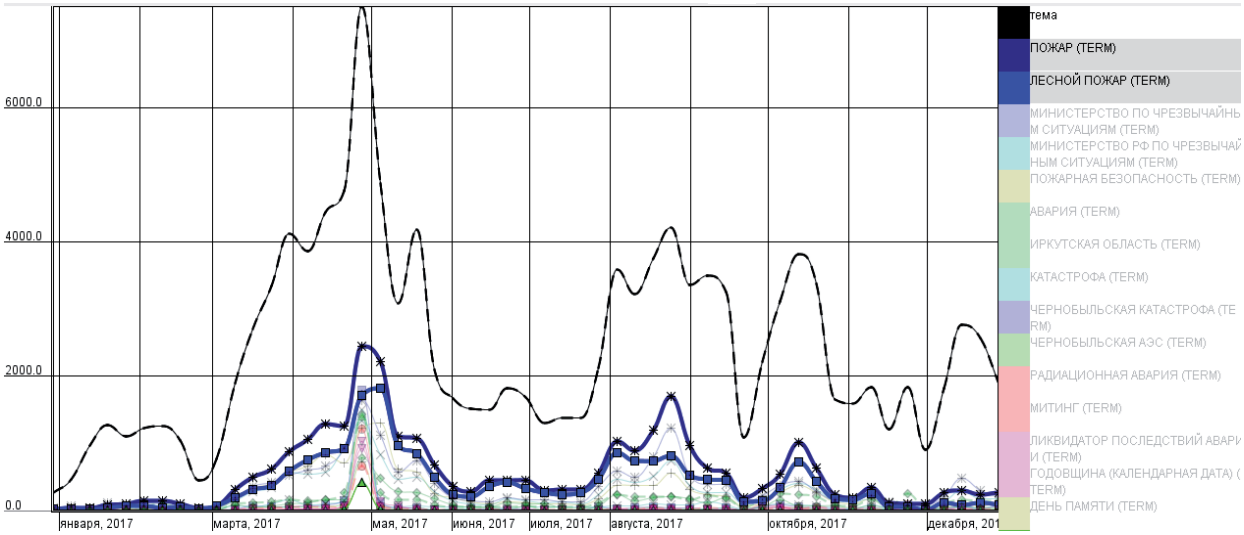


Рисунок 10 - Анализ запроса /Рубрика=«ЭКОЛОГИЧЕСКИЕ УГРОЗЫ»

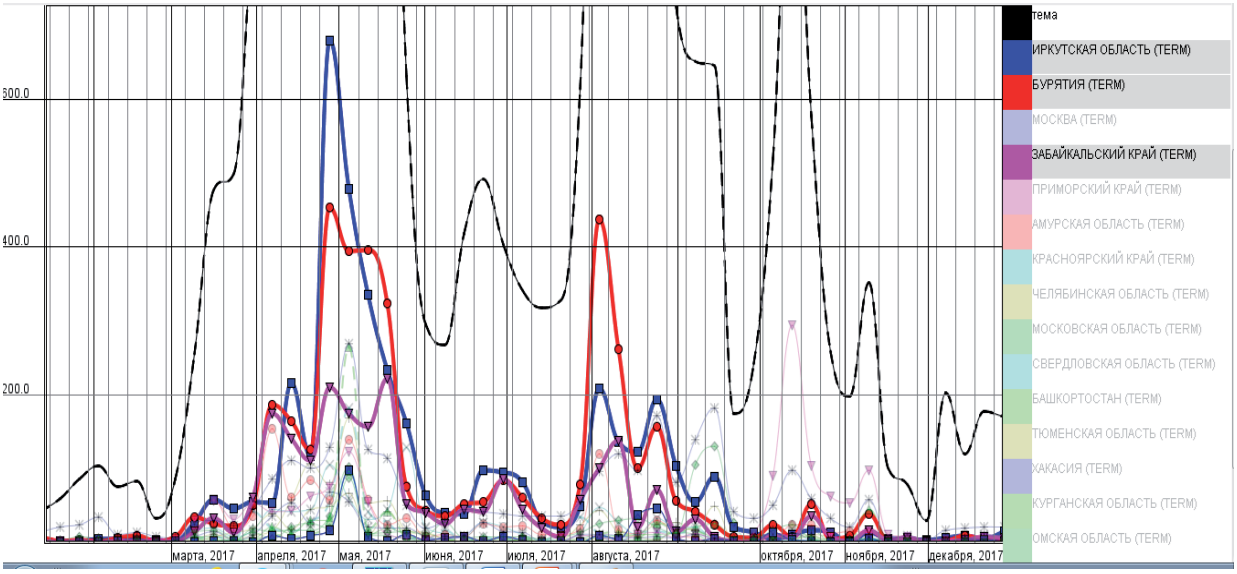


Рисунок 11 - Распределение по субъектам РФ упоминания в новостях 2017 года лесных пожаров

5.4 Когнитивные схемы

Фасеты позволяют выявить основные элементы словарей данных, наиболее характерные для поисковых выборок. Спектрально-фасетный анализ позволяет анализировать вклад в содержимое выборок разных элементов словарей данных в динамике.

Между факторами, определяющими закономерности развития исследуемого процесса или явления могут существовать достаточно сложные зависимости. Для описания сложных зависимостей полезно использовать графовое представление. В среде аналитиков широко известен программный продукт IBM i2[1], предназначенный для исследования взаимосвязей между именованными сущностями. При этом сила связи между сущностями является монотонной функцией от количества документов, в тексте которых упоминаются обе сущности.

[1] https://www.ibm.com/ru-ru/marketplace/analysts-notebook

В рамках данной информационно-аналитической системы используется программное обеспечение визуализации графов GraphView, тесно увязанное с поисковой машиной NearIdx (см. рисунок 12). В NearIdx хранится значительно больше информации, чем простые индексы по именованным сущностям. Соответственно в GraphView граф может включать элементы разных словарей данных.

Граф на основе выборки формируется в NearIdx, при этом имеется возможность выбора различных фасетов, а также любых других сущностей словарей данных, других запросов для отображения на графе. После чего граф разрисовывается и модифицируется в GraphView.

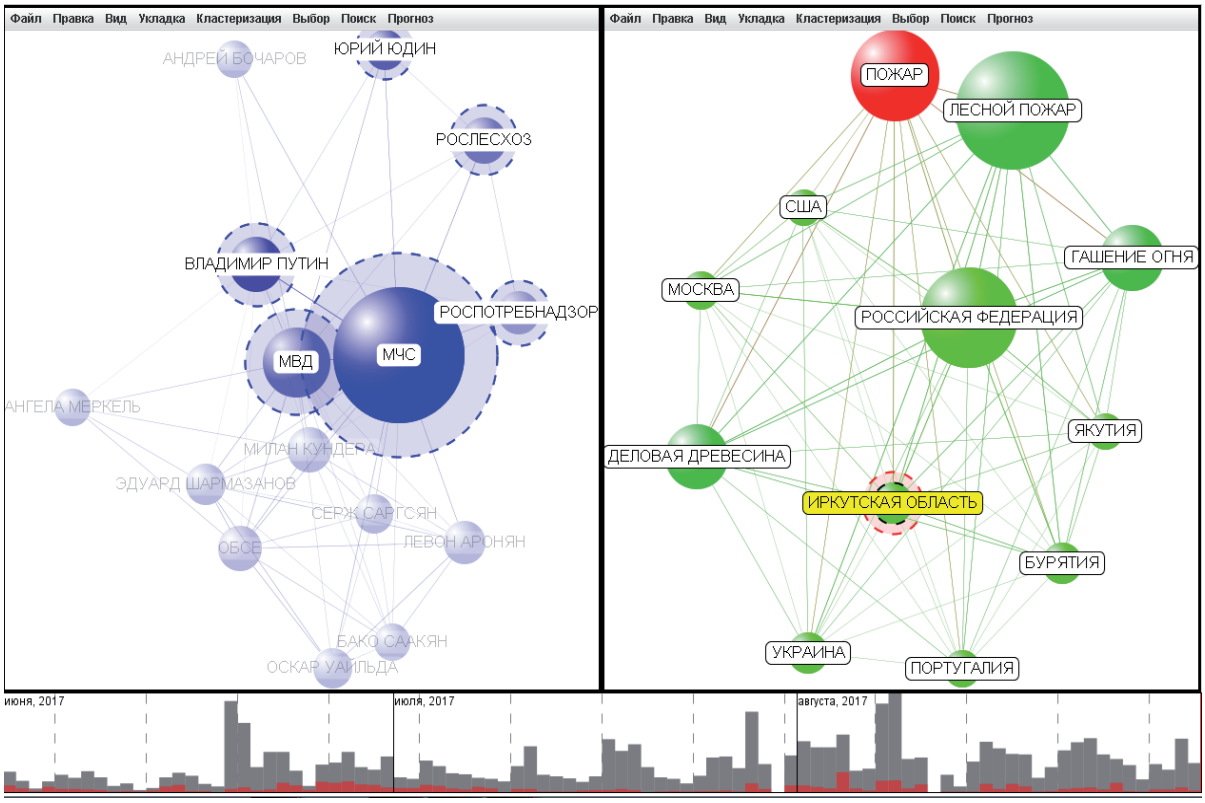


Рисунок 12 - Представление содержания документов по тематике лесных пожаров.
На левой панели представлены упоминаемые персоны и организации, на правой панели – понятия онтологии

GraphView разрабатывался для поддержки труда аналитиков, для которого характерны отработка задач выдвижения гипотез и их быстрой проверки. Так как одной из проблем при аналитическом исследовании графов является наличие большого количества связей между объектами, то в GraphView для лучшего восприятия информации содержимое графа разделено на две панели. Рёбра между объектами на разных панелях не отображаются, но информация о них хранится, и взаимосвязи между объектами на разных панелях визуализируются выделением яркостью цвета между связанными узлами и затемнением несвязанных узлов графа. Таким образом, поддерживается исходный «полный» граф, который виртуально разделяется на несколько отображаемых визуально панелей. При этом узлы графа можно перемещать/копировать с одного подграфа на другой, что полезно, например, для анализа сравнительно небольшого подграфа.

Ниже двух панелей приводится временной ряд количества документов выборки, на котором визуализируется количество документов по выбранным сущностям.

Опыт показывает, что при анализе по именованным сущностям и лингвистическим онтологиям удобно размещать именованные сущности на одном подграфе, а понятия онтологии -

на другом. Тогда один подграф «отвечает» на вопрос «о ком говорилось в текстах выборки», а второй подграф отвечает на вопрос «что говорилось в текстах выборки». При этом выбор одного или нескольких объектов на одном подграфе приводит к обозначению связанных объектов на обоих подграфах на фоне затеняемых несвязанных объектов (рисунок 12).

5.5 Аналитические справки

Изучение стандартных аналитических отчётов, а также отчётов по мониторингу прессы, показывает, что они имеют схожую, достаточно простую структуру. Отчёты посвящены одной теме и, при необходимости, структурируются по частям, относящимся к другим темам. При этом единицей информации являются фрагменты исходных документов. Такая структура может быть смоделирована с использованием качественной поисковой машины.

Действительно, если рассмотреть заглавный запрос Q_0 и множество структурирующих запросов Q_i (будем называть их «подрубрикатором»), тогда общий запрос вида Q_0 & Q_i должен вернуть все релевантные документы (фрагменты). После этого остаётся только избавиться от повторов (например, кластеризовав фрагменты) и расположить фрагменты в соответствии с заданной последовательностью дополнительных запросов.

Рассмотрим в качестве примера запрос по рубрике (рисунок 13):

/УГРОЗЫ= «Демографические угрозы»

В данном автоматически построенном отчёте полученные фрагменты (релевантными считаются фрагменты длиной не более 3 предложений), сгруппированы по сходству и выведены по убыванию даты публикации.

Отчет по запросу:
/УГРОЗЫ="ДЕМОГРАФИЧЕСКИЕ УГРОЗЫ"

Дата отчета: 2017-12-18 20:39:24
Формирование отчета: по документам
Сортировка: по убыванию даты

(0.74) 28.11.2017 22:02:22 **На выплаты первенцам государство потратит за три года 144 млрд рублей [Полит.ру]**

На выплаты первенцам государство потратит за три года 144 млрд рублей На выплаты первенцам государство потратит за три года 144 млрд рублей Младенцы Советник Института современного развития Никита Масленников заявил РИА Новости , что объявленная президентом РФ Владимиром Путиным перезагрузка в демографической политике отвечает чаяниям россиян и поможет исправить ситуацию с рождаемостью в стране. Речь в частности, идет о предложении российского лидера ввести ежемесячные социальные выплаты на первенцев. «В целом эти инициативы абсолютно оправданы, во-первых, потому, что они просто необходимы по жизни, надо думать о том, как мы будем выползать из этой демографической ямы, как стабилизировать численность населения страны на ближайшие 20 лет.

(0.69) 28.11.2017 19:13:05 **Эксперт оценил пользу социальных предложений Путина [РИА Новости]**

Эксперт оценил пользу социальных предложений Путина МОСКВА, 28 ноя — РИА Новости. Инициативы в социальной сфере, озвученные президентом РФ Владимиром Путиным во вторник, помогут России выбраться из демографической "ямы" и ответить на социальные ожидания населения в начале нового политического цикла, считает советник Института современного развития Никита Масленников.
Президент РФ Владимир Путин во вторник озвучил ряд инициатив в социально-демографической сфере.
Тем более что есть там и переходящие деньги на следующий год, и определенный резерв на различного рода меры социальной и иной поддержки заложены и на 2018 год и очевидно на следующий год", — рассказал эксперт. "Поэтому здесь, я думаю, с источниками должно быть все нормально, тем более понятна целевая функция этих мер, потому что мы уже в следующем году сталкиваемся с достаточно серьезной ситуацией на рынке труда, когда примерно 300 тысяч человек с него уйдет и дальше до 2024 года по меньшей мере у нас должны постоянно сокращения с рынка труда, от 300 до 600 тысяч. Поэтому естественно, демографическая проблема требует повышенного внимания с учетом того, что нам нужно стимулировать деторождаемость и рост населения с тем, чтобы на рубеже 2030-2035 годов более-менее выровнять ситуацию", — отметил Масленников.

(0.69) 28.11.2017 11:06:17 **Население Омской области вымирает [КоммерческиеВести]**

Население Омской области вымирает 0 75 С января по октябрь в регионе число умерших было на 2296 человек больше, чем родившихся. В прошлом году отмечался естественный прирост населения на 521 человека. Омскстат обнародовал данные о естественном движении населения региона в январе-октябре 2017 года. **+ 1 фрагм.**

(0.69) 27.11.2017 20:06:25 **Каждый второй нижегородец умирает от болезней системы кровообращения [ИА REGNUM]**

Каждый второй нижегородец умирает от болезней системы кровообращения Эдвард Мунк. У смертного одра. 1893 Нижний Новгород , 27 ноября 2017 . 19:41 — REGNUM Болезни систем кровообращения остаются самыми частыми причинами смерти среди жителей

Рисунок 13 – Отчёт по запросу «Демографические угрозы»

На рисунке 14 представлен пример формы задания параметров построения аналитической справки с заданным подрубрикатором по регионам Сибирского федерального округа. Для задания разделов подрубрикатора указывается порядок, название раздела и запрос, который определяет наполнение данного раздела. На рисунке15 представлен результат построения аналитической справки, структурированной по регионам.

Таким образом, созданная информационно-аналитическая система NearIdx обеспечивает стандартные функции информационного поиска, а также даёт возможность задания запросов и поиска информации с использованием специализированных лексико-терминологических ресурсов (онтологий, рубрикаторов). Аналитический компонент NearIdx предоставляет возможности фасетного анализа, спектрально-фасетного анализа, построения когнитивных схем и аналитических справок, в которых также могут использоваться созданные ресурсы.

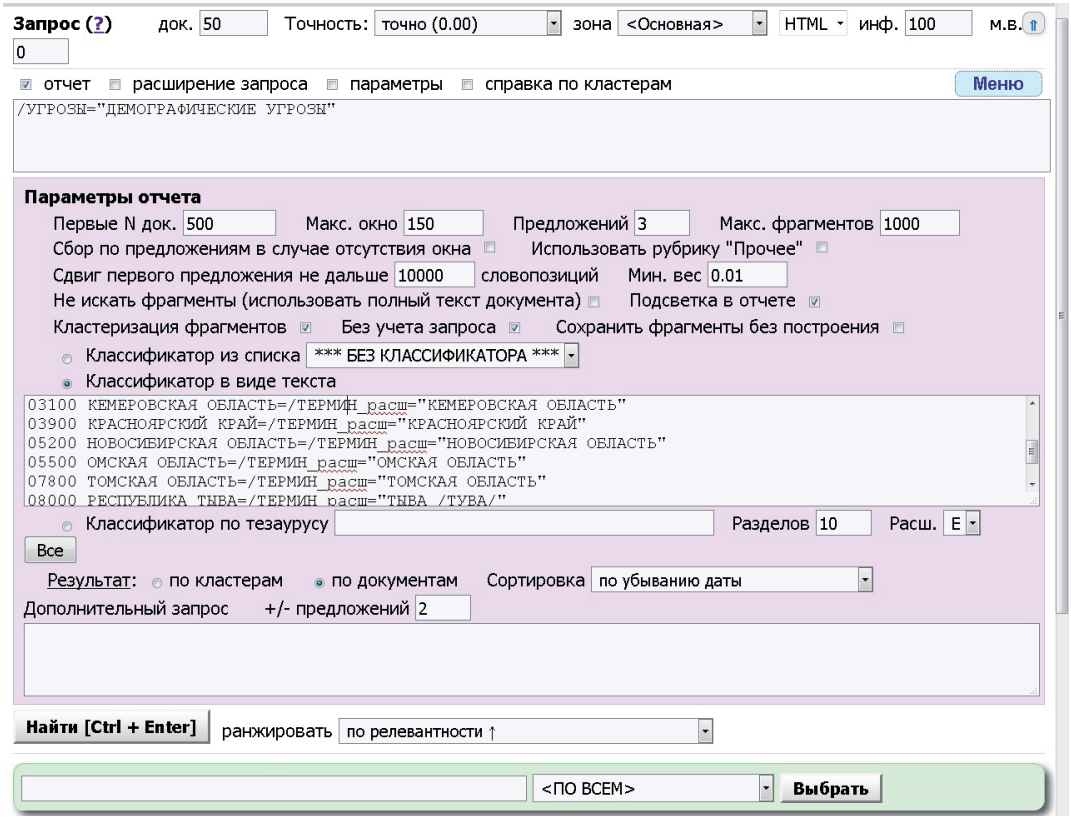


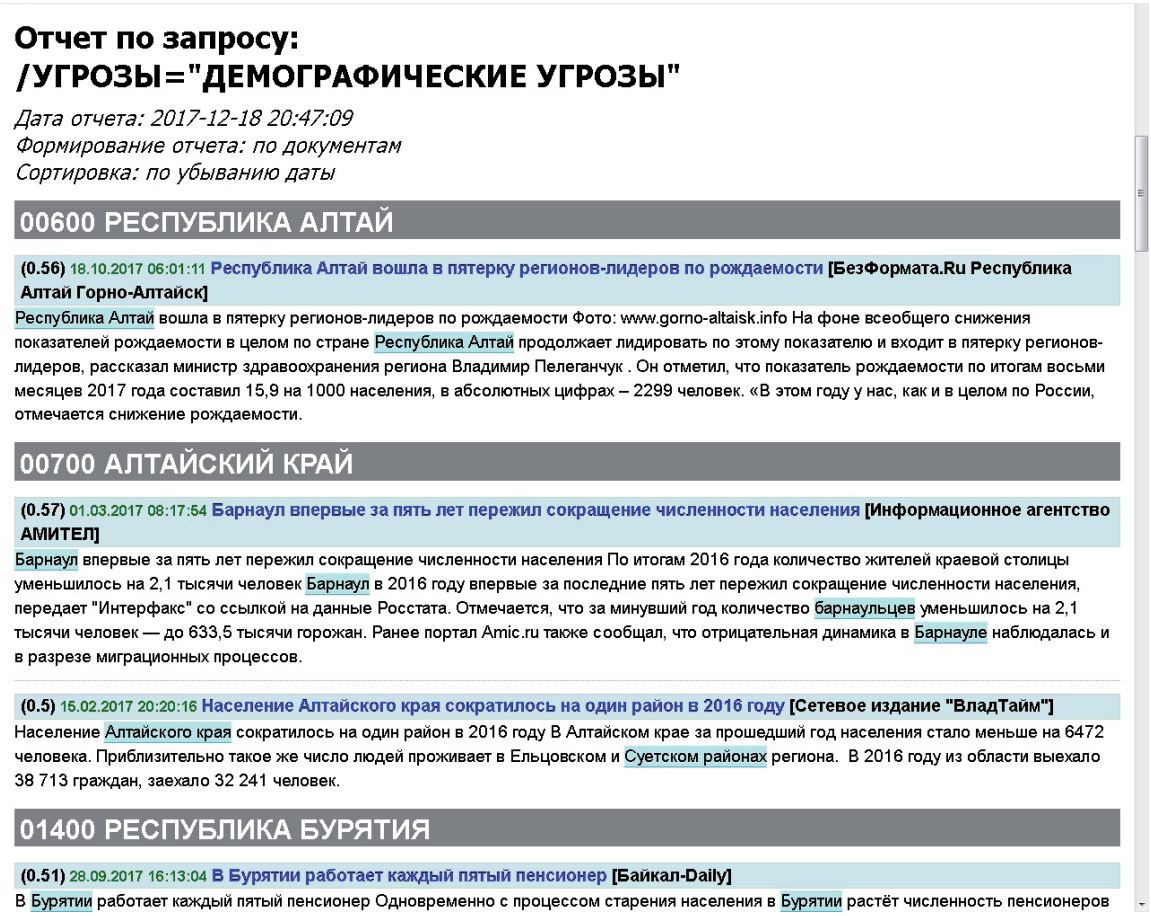
Рисунок 14 – Пример формы задания исходного запроса: рубрика «Демографические угрозы» из рубрикатора угроз (в качестве подрубрик указаны некоторые субъекты Российской Федерации)

Заключение

В статье рассмотрен подход к описанию широкой области национальной безопасности как тезауруса для автоматической обработки документов. Созданный Тезаурус по безопасности имеет модель представления тезауруса РуТез, используется в специализированной информационно-аналитической системе и для автоматической текстовой классификации документов в соответствии с несколькими рубрикаторами, включая рубрикатор угроз, рубрикатор ценностей, рубрикатор региональных проблем и др.

Созданная информационно-аналитическая система NearIdx обеспечивает стандартные функции информационного поиска, а также даёт возможность задания запросов и поиска информации с использованием специализированных ресурсов, включая онтологию и рубри-

каторы. Аналитический компонент NearIdx предоставляет возможности фасетного анализа, спектрально-фасетного анализа, построения когнитивных схем и аналитических справок, в которых также могут использоваться созданные лексико-терминологические ресурсы.



Отчет по запросу:
/УГРОЗЫ="ДЕМОГРАФИЧЕСКИЕ УГРОЗЫ"

Дата отчета: 2017-12-18 20:47:09
Формирование отчета: по документам
Сортировка: по убыванию даты

00600 РЕСПУБЛИКА АЛТАЙ

(0.56) 18.10.2017 06:01:11 **Республика Алтай вошла в пятерку регионов-лидеров по рождаемости [БезФормата.Ru Республика Алтай Горно-Алтайск]**

Республика Алтай вошла в пятерку регионов-лидеров по рождаемости Фото: www.gorno-altaisk.info На фоне всеобщего снижения показателей рождаемости в целом по стране Республика Алтай продолжает лидировать по этому показателю и входит в пятерку регионов-лидеров, рассказал министр здравоохранения региона Владимир Пелеганчук . Он отметил, что показатель рождаемости по итогам восьми месяцев 2017 года составил 15,9 на 1000 населения, в абсолютных цифрах – 2299 человек. «В этом году у нас, как и в целом по России, отмечается снижение рождаемости.

00700 АЛТАЙСКИЙ КРАЙ

(0.57) 01.03.2017 08:17:54 **Барнаул впервые за пять лет пережил сокращение численности населения [Информационное агентство АМИТЕЛ]**

Барнаул впервые за пять лет пережил сокращение численности населения По итогам 2016 года количество жителей краевой столицы уменьшилось на 2,1 тысячи человек Барнаул в 2016 году впервые за последние пять лет пережил сокращение численности населения, передает "Интерфакс" со ссылкой на данные Росстата. Отмечается, что за минувший год количество барнаульцев уменьшилось на 2,1 тысячи человек — до 633,5 тысячи горожан. Ранее портал Amic.ru также сообщал, что отрицательная динамика в Барнауле наблюдалась и в разрезе миграционных процессов.

(0.5) 15.02.2017 20:20:16 **Население Алтайского края сократилось на один район в 2016 году [Сетевое издание "ВладТайм"]**

Население Алтайского края сократилось на один район в 2016 году В Алтайском крае за прошедший год населения стало меньше на 6472 человека. Приблизительно такое же число людей проживает в Ельцовском и Суетском районах региона. В 2016 году из области выехало 38 713 граждан, заехало 32 241 человек.

01400 РЕСПУБЛИКА БУРЯТИЯ

(0.51) 28.09.2017 16:13:04 **В Бурятии работает каждый пятый пенсионер [Байкал-Daily]**

В Бурятии работает каждый пятый пенсионер Одновременно с процессом старения населения в Бурятии растёт численность пенсионеров

Рисунок 15 - Отчет по запросу «Демографические угрозы» по нескольким регионам России

Благодарности

Работа частично поддержана грантом РФФИ (проект 16-29-09606) и контрактом Министерства образования и науки Российской Федерации (№ 14.601.21.0018).

Список источников

[1] ***Abbas, A., Zhang, L., Khan, S.*** (2014). A literature review on the state-of-the-art in patent analysis / *A. Abbas, L., Zhang, L., S. Khan. //* World Patent Information. – 2014. – V. 37. – P.3-13.

[2] ***Efimenko, I.*** Peaks, Slopes, Canyons and Plateaus: Identifying Technology Trends Throughout the Life Cycle / *I. Efimenko, V Khoroshevsky* // International Journal of Innovation and Technology Management. – 2017. – 14(02).

[3] ***Ena, O.*** A methodology for technology trend monitoring: the case of semantic technologies / *O. Ena, N. Mikova, O, Saritas, A. Sokolova* // Scientometrics. – 2016. – 108 (3). – P.1013-1041.

[4] ***Nassirtoussi, A.K***. Text mining for market prediction: A systematic review / *A.K. Nassirtoussi, S, Aghabozorgi, T.Y. Wah, D.C, Ngo.* // Expert Systems with Applications. – 2014. – V. 41 (16). – P.7653-7670.

[5] ***Трошин, Д.В.*** Основы концептуальной модели источников угроз экономической безопасности на национальном уровне / Д.В. Трошин // Онтология проектирования. – 2017. – Т. 7, № 4(26). – С.410-422. – DOI: 10.18287/2223-9537-2017-7-4-410-422.

[6] ***Лукашевич, Н.В.*** Тезаурусы в задачах информационного поиска / Н.В. Лукашевич. – М.: Издательство МГУ, 2011.
[7] ***Shane, S.*** Isis displaying a deft command of varied media / S. *Shane, B.Hubbard* // New York Times 31. – 2014.
[8] ***Weimann, G***. New Terrorism and New Media / G. *Weimann* // Wilson Center Common Labs. – 2014.
[9] ***Kohlmann, E.*** Profiles of foreign fighters in Syria and Iraq / *E. Kohlmann, L. Alkhouri* // CTC Sentinel, September. - 2014. - V. 29.
[10] ***Finlayson, M.A.*** The N2 corpus: A semantically annotated collection of Islamist extremist stories / *M.A. Finlayson, J.R.Halverson, S.R. Corma,* // In LREC-2014. – 2014. – P.896-902.
[11] ***Sela, S.*** Changes in the discourse of online hate blogs: The effect of Barack Obama's election in 2008 */ S. Sela, T. Kuflik, G.S. Mesch.* // First Monday. – 2012. – V.17, N.11.
[12] ***Kwok, I.*** 2013. Locate the hate: Detecting tweets against blacks / *I. Kwok, Y. Wang //* In AAAI-2013. – 2013.
[13] ***Zeerak, W***. Hateful symbols or hateful people? predictive features for hate speech detection on twitter / *W. Zeerak, D. Hovy* // Proceedings of NAACL-HLT-2016. – 2016. – P.88-93.
[14] **Nobata, Ch.** Abusive language detection in online user content / *Ch. Nobata, J. Tetreault, A. Thomas, M. Yashar, Yi Chang* //, in Proc. of the 25th International Conference on World Wide Web. – 2016. – P.145-153.
[15] ***Schmidt, A***. A survey on hate speech detection using natural language processing / *A. Schmidt, M. Wiegand* // SocialNLP-2017. – 2017.
[16] ***Lim, S.*** MalwareTextDB: A Database for Annotated Malware Articles / S. *Lim, A. Muis, W. Lu, C. Ong* // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. – 2017. – V.1.
[17] ***Kirillov, I.*** 2010. Malware Attribute Enumeration and Characterization / *I. Kirillov, D, Beck, P, Chase, R. Martin* // The MITRE Corporation, Tech. Rep. – 2010.
[18] ***Gorokhov, O.*** Convolutional Neural Networks for Unsupervised Anomaly Detection in Text Data / *O. Gorokhov, M. Petrovskiy, I. Mashechkin* // International Conference on Intelligent Data Engineering and Automated Learning. – Springer, Cham, 2017. – P.500-507.
[19] ***O'brien, S.P.*** Crisis early warning and decision support: Contemporary approaches and thoughts on future research */ S.P. O'brien.* // International studies review. – 2010. – 12(1). – P.87-104.
[20] ***Wang, W.*** Growing pains for global monitoring of societal events. / *W. Wang, R. Kennedy, D. Lazer, N. Ramakrishnan* // Science. – 2016. – 353(6307). – P.1502-1503.
[21] ***Лукашевич, Н.В***. Проектирование лингвистических онтологий для информационных систем в широких предметных областях / Н.В. Лукашевич, Б.В. Добров // Онтология проектирования. – 2015. – Т. 5. – №. 1 (15) – С.47-69.
[22] ***Loukachevitch, N.*** RuThes linguistic ontology vs. Russian wordnets / *N. Loukachevitch, B. Dobrov* // Proceedings of Global WordNet Conference GWC-2014. 2014.
[23] ***Loukachevitch, N.*** RuThes-Lite, a publicly available version of Thesaurus of Russian language RuThes / *N. Loukachevitch, B. Dobrov, I. Chetviorkin* // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference в Dialogue-2014. – 2014. – P.340-349.
[24] ***Loukachevitch, N.*** The Sociopolitical Thesaurus as a resource for automatic document processing in Russian / *N. Loukachevitch, B. Dobrov* // Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication. – 2015. – V. 21, N.2. – P.237-262.
[25] ***Dobrov, B.*** Development of Linguistic Ontology on Natural Sciences and Technology / B. Dobrov, N. Loukachevitch // Proceedings of Linguistic Resources and Evaluation Conference LREC-2006. – 2006.

ONTOLOGICAL RESOURCES AND INFORMATION-ANALYTICAL SYSTEM IN SECURITY DOMAIN

N.V. Loukachevitch[1], B.V. Dobrov[2], A.M. Pavlov[3], S.V. Shternov[4]

Research Computing Center of Lomonosov Moscow State University, Moscow, Russia
[1]louk_nat@mail.ru, [2]dobrov_bv@mail.ru, [3]pavlov.andrew.m@gmail.com, [4]shternov@gmail.com

Abstract

The article considers an approach to describing a broad area of national security as a thesaurus for automatic document processing. The created thesaurus of safety has a model of representation of the thesaurus RuThes. The security thesaurus is used in a specialized information and analytical system and for automatic text classification of documents in ac-

cordance with several systems of subject headings, including the Threat categories, the Value, Regional problem categories, etc. The created Information and Analytical System NearIdx provides standard information search functions, and it also the ability to specify queries and search for information using specialized resources, including ontology and subject headings. The analytical component NearIdx provides the possibilities of facet analysis, spectral-facet analysis, construction of cognitive schemes and analytical references, in which the ontological resources can also be used.

Key words: *thesaurus, ontology, national security, information retrieval, subject headings, automatic text categorization.*

Citation: *Loukachevitch NV, Dobrov BV, Pavlov AM, Shternov SV.* Ontological Resources and Information-Analytical System in Security Domain [In Russian]. *Ontology of designing*. 2018; 8(1): 74-95. - DOI: 10.18287/2223-9537-2018-8-1-74-95.

Acknowledgment

The work was partially supported by the RFBR grant (project 16-29-09606) and the contract of the Ministry of Education and Science of the Russian Federation (No. 14.601.21.0018).

References

[1] ***Abbas A, Zhang L, Khan S.*** A literature review on the state-of-the-art in patent analysis. World Patent Information; 2014; 37: 3-13.
[2] ***Efimenko I, Khoroshevsky V***. Peaks, Slopes, Canyons and Plateaus: Identifying Technology Trends Throughout the Life Cycle. International Journal of Innovation and Technology Management; 2017; 14(02).
[3] ***Ena O, Mikova N, Saritas O, Sokolova A.*** A methodology for technology trend monitoring: the case of semantic technologies. Scientometrics; 2016; 108(3); 1013-1041.
[4] ***Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DC***. Text mining for market prediction: A systematic review. Expert Systems with Applications; 2014; 41(16): 7653-7670.
[5] ***Troshin DV***. Foundations of a conceptual model of economic security threat sources at the national level [In Russian]. *Ontology of designing*. 2017; 7(4): 410-422. - DOI: 10.18287/2223-9537-2017-7-4-410-422.
[6] ***Lukashevich NV.*** Thesauri in information-retrieval tasks [in Russian]. – M.: MSU, 2011.
[7] ***Shane S, Hubbard B***. Isis displaying a deft command of varied media. New York Times; 2014; 31.
[8] ***Weimann G***. New Terrorism and New Media. Wilson Center Common Labs; 2014.
[9] ***Kohlmann E, Alkhouri L.*** Profiles of foreign fighters in Syria and Iraq. CTC Sentinel, September; 2014; 29.
[10] ***Finlayson MA, Halverson JR, Corman SR.*** The N2 corpus: A semantically annotated collection of Islamist extremist stories. In LREC-2014; 2014: 896-902.
[11] ***Sela S, Kuflik T, Mesch GS***. Changes in the discourse of online hate blogs: The effect of Barack Obama's election in 2008 // First Monday; 2012; 17 (11).
[12] ***Kwok I, Wang Y***. Locate the hate: Detecting tweets against blacks. In AAAI-2013; 2013.
[13] ***Zeerak W, Hovy D***. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. Proceedings of NAACL-HLT-2016; 2016: 88-93.
[14] ***Nobata Ch, Tetreault J, Thomas A, Yashar M, Chang Yi***. Abusive language detection in online user content, in Proc. of the 25th International Conference on World Wide Web; 2016: 145–153.
[15] ***Schmidt A, Wiegand M.*** A survey on hate speech detection using natural language processing. Proceedings of SocialNLP; 2017.
[16] ***Lim S, Muis A, Lu W, Ong C.*** MalwareTextDB: A Database for Annotated Malware Articles. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1; 2017.
[17] ***Kirillov I, Beck D, Chase P, Martin R***. Malware Attribute Enumeration and Characterization. The MITRE Corporation, Tech. Rep.; 2010.
[18] ***Gorokhov O, Petrovskiy M, Mashechkin I.*** Convolutional Neural Networks for Unsupervised Anomaly Detection in Text Data. International Conference on Intelligent Data Engineering and Automated Learning. Springer, Cham; 2017: 500-507.
[19] ***O'brien SP***. Crisis early warning and decision support: Contemporary approaches and thoughts on future research // International studies review; 2010; 12(1): 87-104.
[20] ***Wang W, Kennedy R, Lazer D, Ramakrishnan N.*** Growing pains for global monitoring of societal events. Science; 2016; 353(6307): 1502-1503.

[21] ***Lukashevich NV, Dobrov BV***. Designing linguistic ontologies for information systems in broad domains [in Russian]. Ontology of designing. 2015; 5(1): 47-69.
[22] ***Loukachevitch N, Dobrov B.*** RuThes linguistic ontology vs. Russian wordnets. Proceedings of Global WordNet Conference GWC-2014; 2014.
[23] ***Loukachevitch N, Dobrov B, Chetviorkin I.*** RuThes-Lite, a publicly available version of Thesaurus of Russian language RuThes. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference в Dialogue, Bekasovo, Russia; 2014: 340-349.
[24] ***Loukachevitch N, Dobrov B.*** The Sociopolitical Thesaurus as a resource for automatic document processing in Russian. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication; 2015; 21 (2): 237-262.
[25] ***Dobrov B, Loukachevitch N.*** Development of Linguistic Ontology on Natural Sciences and Technology. Proceedings of Linguistic Resources and Evaluation Conference LREC-2006. – 2006.

Сведения об авторах

Лукашевич Наталья Валентиновна, 1964 г. рождения. Окончила факультет вычислительной математики и кибернетики МГУ имени М.В. Ломоносова в 1986 г., к.ф.-м.н. (1989), д.т.н. (2016). Ведущий научный сотрудник НИВЦ МГУ имени М.В. Ломоносова. В списке научных трудов более 180 работ в области автоматической обработки текстов, представления знаний.

Natalia Valentinovna Loukachevitch (b.1964) graduated from Lomonosov Moscow State University in 1986, PhD (1989), Doctor of Sciences (2016). She is a leading researcher in Research Computing Center of Lomonosov Moscow State University. She is an author of more 180 scientific papers in the field of natural language processing, knowledge representation.

Добров Борис Викторович, 1963 г. рождения. Окончил факультет вычислительной математики и кибернетики МГУ имени М.В. Ломоносова в 1985 г., к.ф.-м.н. (1988). Заведующий лабораторией НИВЦ МГУ имени М.В. Ломоносова. В списке научных трудов более 120 работ в области информационного поиска, онтологий

Boris Vikotorovich Dobrov (b. 1963) graduated from Lomonosov Moscow State University in 1985, PhD (1988). Chief of laboratory in Research Computing Center of Lomonosov Moscow State University. He is an author of more than 120 publications in the field of information retrieval, ontologies.

Павлов Андрей Михайлович**,** 1986 г. рождения. Окончил факультет вычислительной математики и кибернетики МГУ имени М.В. Ломоносова в 2009 г. Программист НИВЦ МГУ имени М.В. Ломоносова. Имеет публикации по тематике кластеризации новостного потока.

Andrew Mikhailovich Pavlov (b. 1986) graduated from Lomonosov Moscow State University in 2009. Programmer of Research Computing Center of Lomonosov Moscow State University. He is an author of publications in the field of news clusterization.

Штернов Сергей Владимирович**,** 1980 г. рождения. Окончил Московский авиационный институт в 2003 г. Программист НИВЦ МГУ имени М.В. Ломоносова. Имеет публикации по тематике информационного поиска и автоматической обработки текстов.

Sergey Vladimirovich Shternov (b. 1980) graduated from Moscow Aviation Institute in 2003. Programmer of Research Computing Center of Lomonosov Moscow State University. He is an author of publications in the fields of information retrieval, natural language processing.