Методы автоматического построения формализованного представления содержания материалов электронных средств массовых коммуникаций для решения задачи мониторинга и оценки деятельности органов власти

© Ю.В. Никитин © Ал-др А. Хорошилов © Ал-ей А. Хорошилов
Институт проблем информатики Российской академии наук (ИПИ РАН),
Москва
yuri.v.nikitin@gmail.com khoroshilov@mail.ru alex_khoroshilov@mail.ru

Аннотация

В данной статье рассматриваются возможности создания формализованного представления информационных публикаций в сети Интернет для получения показателей количественной оценки деятельности органов власти по материалам таких публикаций. Также рассматриваются методы построения формализованного описания информационных сообщений и методы адаптации автоматизированных средств семантической обработки сообщений для получения наиболее адекватных результатов анализа в заданной предметной области.

1 Введение

В настоящее время приобретает все большую актуальность получение обратной связи от населения при оценке деятельности органов власти. Такая оценка востребована как самими органами власти для принятия оперативных решений, их вышестоящими и надзорными структурами, так и различными общественными и исследовательскими организациями по мониторингу общественного мнения.

Одним из возможных источников сбора информации для проведения подобных исследований является информационное пространство электронных средств массовых коммуникаций в сети Интернет, включающее такие источники информации, как электронные средства массовой информации (СМИ), публикации в блогах и на форумах, сообщения в социальных сетях, сервисы коротких сообщений (например, Twitter) и электронные ресурсы обратной связи с населением на государственных порталах по приему жалоб и обращений граждан.

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13-16 октября 2014 г.

Все более широкое приобщение населения к электронным информационным ресурсам, в том числе рост популярности социальных сервисов среди молодежи и стимулирование государством использования гражданами электронных госуслуг, с одной стороны, предоставляет все больше возможностей для оперирования данными, получаемыми в электронном виде по сети Интернет, с другой стороны, все более возрастающий объем подобной информации с каждым годом в значительной мере усложняет задачу обработки этих сведений, затрудняя деятельность экспертов-аналитиков в области оценки общественного мнения о деятельности органов власти.

В этих условиях эксперты-аналитики все чаще ставят задачи перехода от качественных экспертных оценок по результатам изучения материалов информационных сообщений к автоматизированным количественным методам оценки общественного мнения о деятельности органов власти. Такие методы имеют два неоспоримых взаимосвязанных преимущества: во-первых, дают возможность получения более объективных статистических показателей, благодаря росту объема обрабатываемых данных, во-вторых, максимально освобождают результаты оценки от субъективного экспертного представления о действительной ситуации по тем же самым причинам и вследствие более глубокой автоматизации обработки этого объема информации, воспринять которую в ее исходном виде за короткий период времени ни один эксперт просто не в состоянии.

В то же время решение данной задачи сталкивается с рядом комплексных проблем. Прежде всего, необходимы современные и адекватные средства семантического анализа неструктурированной текстовой информации. Далее необходимо определить набор показателей для проведения количественного статистического анализа. В настоящее время имеется серьезная проблема с отсутствием определенного набора показателей для подобных оценок. Это связано с тем, что ранее у занимающихся этими вопросами исследователей не было большого опыта обработки

данных такого объема, а также не было четкого представления о возможностях формализации текстов, доступных методах преобразования неструктурированных данных в структурированные – т.е. не было понятия о доступном инструментарии.

В данной статье мы предлагаем возможный инструментарий для проведения таких оценок и предлагаем методы автоматической обработки неструктурированной текстовой информации, позволяющей разработать модель количественных показателей. Мы также рассматриваем условия, при которых приведенные методы будут адекватны поставленной задаче в рамках заданной предметной области, посредством настройки декларативных средств к заданной предметной области с учетом большого объема обрабатываемых данных.

Мы провели моделирование процесса автоматизированной обработки текстов с целью получения количественных показателей на примере информационных сообщений Интернет-СМИ и пользователей социальной сети «ВКонтакте» о деятельности органов власти Ханты-Мансийского автономного округа (ХМАО).

2 Программно-техническое обеспечение

2.1 Общие требования к процессу автоматизации

Для автоматизации решения задачи оценки деятельности органов власти по материалам Интернет-публикаций необходимо обеспечить программно-техническую реализацию следующих основных ее подзадач:

1. **Мониторинг** материалов Интернет-публикаций:

а) консолидация информационных потоков различных типов: ленты новостей на сайтах и порталах Интернет-СМИ, органов власти и комментарии пользователей к ним, публикации на форумах и в блогах, сообщения пользователей социальных сетей и сервисов коротких сообщений, электронные обращения граждан и другие доступные ресурсы;

б) оперативный мониторинг изменения информации на подключенных ресурсах (информационных источниках) и сбор (считывание и загрузка) содержимого текстовых материалов;

в) унификация форматов представления текстовых сообщений, извлечение возможных реквизитов публикаций, аналогичных библиографическому описанию (источник, рубрика, автор, наименование публикации, временной период и т.п.), в зависимости от типа источника.

2. **Лингвистическая обработка** неструктурированных текстов:

а) автоматическое создание формализованного представления смысловой структуры текста;

б) кластеризация сходных по смысловому содержанию текстов – группировка текстов публикаций по темам (информационным поводам);

в) выделение и классификация объектов, их признаков и отношений между ними и классификация текстов по типам отношений автора сообщения к основным объектам мониторинга;

г) автоматизированная настройка декларативных (словарных) средств лингвистического процессора на заданную предметную область.

3. **Обработка данных с применением технологий «Big Data»**:

а) обеспечение распределенной массово-параллельной лингвистической и статистической обработки загружаемых данных;

б) обеспечение масштабируемости на множество узлов обработки без деградации инфраструктуры обработки данных [2, 3].

Подзадача 1 (мониторинг материалов) является достаточно тривиальной задачей, имеющей множество программно-технических решений. Подзадача 3 (обработка данных с применением технологий «Big Data») является самостоятельным направлением исследований, и мы оставляем ее за рамками данной статьи, при этом учитывая требования по распределенности и масштабируемости к средствам лингвистической обработки текстов.

В данной статье мы рассматриваем подзадачу 2 (лингвистическая обработка неструктурированных текстов), методы ее обеспечения и требования к созданию декларативных (словарных) средств [1, 7].

2.2 Требования к лингвистическому программному обеспечению

Современные системы автоматизированной семантической обработки неструктурированной текстовой информации, разрабатываемые для решения задач данного типа, должны обеспечивать выполнение следующих процедур лингвистического анализа текстов [1–9]:

а) графематический анализ текста;

б) морфологический анализ слов;

в) семантико-синтаксический анализ текстов;

г) концептуальный анализ текстов;

д) дистрибутивно-статистический анализ текстов.

Графематический анализ предназначен для предварительного анализа текста по представляющей его последовательности символов [2, 3]. В результате этого анализа определяется язык текста, устанавливаются местоположения слов, предложений, абзацев, фамильно-именной группы, дат, адресов и т.п.

Морфологический анализ слов естественных языков предназначен для определения структуры слов и назначения им грамматических признаков, необходимых для выполнения последующих процедур автоматической обработки текстовой информации (например, синтаксического и концептуального анализа текстов) [8, 9].

Семантико-синтаксический анализ текстов проводится с целью формализованного представления их структуры – выделения в них смысловых единиц и установления связей между ними. При этом структура текстов может интерпретироваться по-разному и описываться на различных формализованных языках [7, 8].

Концептуальный анализ текстов предназначен для определения смысловой структуры текстов, выявления их понятийного (концептуального) состава текстов и установления связей между наименованиями понятий [7, 8].

Дистрибутивно-статистический анализ текстов естественных языков предназначен для установления статистических закономерностей совместной встречаемости наименований понятий [1, 9].

2.3 Требования к извлечению данных

Основной задачей при выполнении семантической обработки неструктурированной текстовой информации является представление смысловой структуры текста в формализованном виде [8, 9].

В классическом виде формализованное представление текстового содержания документа должно содержать:

а) библиографические реквизиты (например, информационный источник, рубрика, автор, наименование и дата публикации и т.п.);

б) аннотацию или реферат документа;

в) список ключевых выражений;

г) классификацию документа по смысловому содержанию – отнесение его к той или иной рубрике и кластеризация (группировка) текстов публикаций по темам (информационным поводам).

Рассмотрим более подробно каждый из реквизитов формализованного описания применительно к поставленной задаче.

Библиографические реквизиты выделяются на этапе мониторинга информационных публикаций. Структура реквизитов документа (статьи, сообщения, комментария) закладывается в шаблон загрузки и парсинга (разбора на структурные элементы) страницы с текстом публикации для каждого конкретного информационного источника. Соответственно, выделение данных реквизитов происходит на стадии унификации формата документа.

При этом на уровне данных реквизитов, определенных еще до проведения лингвистического анализа, проводится классификация документа, основанная на типе информационного источника. Такая классификация может проводиться по различным основаниям:

а) тип документа (например, статья в СМИ, официальное сообщение, публикация в блоге, комментарий в социальной сети, электронное обращение граждан и т.п.);

б) степень доверия (например, официальный источник, аккредитованное СМИ, «бульварная пресса», подписанное обращение гражданина, анонимное сообщение и т.п.);

в) вид сообщения (например, информационное сообщение, жалоба, комментарий к сообщению, мнение пользователей сети и т.п.);

г) отношение к власти (например, подведомственный источник, аффилированный источник, оппозиционный источник, независимая пресса и т.п.).

Аннотация к тексту документа в общем случае, может быть авторской и изначально сопровождать данный текст, а может отсутствовать в исходном тексте, тогда средствами лингвистического процессора создается автоматический реферат документа [9].

В отличие от автореферата (авторского реферата, отражающего авторское представление о важных, на его взгляд, тезизах документа), автоматический реферат выделяет значимое содержимое дистрибутивно-статистическим способом, основываясь на реальных смысловых акцентах в тексте и на значимости терминов (объектов) в заданной предметной области.

Так же в отличие от автореферата, который может представлять собой стандартный шаблон с аналитическими ответами на ключевые вопросы о содержании документов данного типа, т.е. фактически являющимся пересказом (например, автореферат диссертации), автоматический реферат содержит реальный, не измененный текст документа.

Зачастую документ, содержащийся в базе данных (БД), может интересовать пользователей в разрезе различных тематик. В таком случае необходимо подготовить автоматический контекстный реферат документа в разрезе рассматриваемой тематики или поискового запроса.

В отличие от аннотации такой реферат необходимо строить каждый раз заново с учетом потребности пользователя в той или иной информации, а также ограничений на объем реферата.

С помощью автоматического реферирования также можно выравнивать объемы сравниваемых по смыслу документов – для задач кластеризации документов с аналогичным содержанием.

Ключевые выражения применительно к материалам электронных публикаций в рамках нашей задачи не являются важными средством визуализации смыслового содержания текста, в отличие, например, от текстов научно-технических публикаций [4–8].

При этом список ключевых выражений (наиболее значимых для данного текста с учетом предметной области) и выделенных из текстов объектов (например, персоны, должности, должностные лица, организации, территории,

производственные объекты, географические объекты, бренды) играет ключевую роль в построении формализованного описания документа для его последующего семантического анализа [2, 3].

Как и в случае с аннотацией, в отличие от авторских ключевых выражений (указанных вручную и отражающих авторское представление о важных, на его взгляд, терминах документа), данные выражения выделяются семантико-синтаксическим и словарным методами, основываясь на реальном содержании текста и на значимости терминов в заданной предметной области (выявленной дистрибутивно-статистическим методом).

Выделение объектов из списка ключевых выражений, таких как должностные лица, организации, территории, производственные объекты, географические объекты и бренды, представляет отдельный интерес, т.к. данные выражения зачастую являются объектами мониторинга.

В отличие от обычных ключевых выражений, чаще предназначенных только для визуализации смыслового содержания, и чуть реже – фигурирующих в качестве тематических тегов документа, выделение объектов мониторинга из текстов позволяет определить основные опорные точки для формализованных показателей:

1) выделить из текста объекты и их предикаты;

2) дать классификацию объектов и отношений к объектам на основе классификации предикатов;

3) установить по тексту связи между объектами.

Классификация объектов проводится на основе базового классификатора лингвистического процессора и дополнительного специально созданного по корпусу текстов классификатора заданной предметной области.

Базовый классификатор содержит только основные классы (например, персона, должность, географический объект, дата-время, обычный концепт (термин) и т.п.)

Классификатор предметной области позволяет задать более конкретную классификацию применительно к заданной предметной области (например, орган федеральной, региональной, муниципальной власти, его подразделение или подведомственное предприятие и т.п.).

Классификация объектов позволяет отнести информационное сообщение к той или иной рубрике, а также применительно к нашей задаче определить к какой ветви структуры органов власти относится объект мониторинга для более нацеленного анализа высказываний, приведенных в публикациях.

Классификация предикатов позволяет установить отношения к основным объектам мониторинга (органам власти и государственным функциям) авторов суждений, приведенных в информационных сообщениях – позитивные, негативные или нейтральные.

Кластеризация (группировка) документов выполняется на последнем этапе лингвистической обработки документов.

3 Экспериментальные данные

3.1 Предметная область и исходные данные для испытаний

Для проверки изложенной гипотезы авторы провели моделирование процесса автоматизированной обработки текстов с целью получения количественных показателей на примере информационных сообщений Интернет-СМИ и пользователей социальной сети «Вконтакте» о деятельности органов власти Ханты-Мансийского автономного округа (ХМАО).

Сначала мы провели сбор и анализ региональных публикаций ХМАО с целью автоматизированного выделения наименований органов власти, объектов инфраструктуры и должностных лиц в объеме около 1000 статей по материалам Интернет-СМИ ХМАО, сайтов органов государственной власти ХМАО (www.admhmao.ru) и городского портала органов местного самоуправления Ханты-Мансийска (www.admhmansy.ru).

Далее мы собрали региональные пользовательские публикации социальной сети «ВКонтакте» в объеме около 650 Мб неформатированного текста для анализа лексики, определяющей тональности высказываний граждан.

Кроме того нами был проанализирован представительный набор коротких сообщений Twitter, который мы рассматривали только для анализа и сравнения лексического состава сообщений информационных источников различных типов, и не использовали в дальнейшем для проведения нашего эксперимента по извлечению данных.

По результатам первичного анализа текстов сделаны выводы о необходимости разделения настройки декларативных средств по трем отдельным корпусам текстов с учетом их особенностей:

1) официальные тексты, публикации СМИ, формальные заявления, обращения граждан, экспертные заключения – любые связные тексты со строгим языковым стилем;

2) тексты социальных сетей, блогосферы и форумов – характеризуются большим количеством орфографических ошибок, опечаток, неправильно употребляемых значений слов, сокращений, жаргонизмов и профессионализмов, неологизмов, молодежной неформальной, нецензурной лексики, несвязной структурой текстов и анафорическими связями с предыдущими комментариями и сообщениями;

3) короткие сообщения Twitter (твиты) – SMS-подобные текстовые сообщения строго ограниченного размера, характеризующиеся тезисным стилем изложения информации, большим количеством сокращений и наличием хэш-тегов вида #subject, при этом часть из этих тегов фигурирует в качестве дополнительных идентификаторов темы, расположенных в начале или в конце сообщения, а часть из низ являются непосредственными членами синтаксической структуры предложения (т.е. значимыми словами в связном тексте).

3.2 Использование лингвистического программного обеспечения

Для проведения экспериментов мы использовали разработанное авторами статьи лингвистическое программное обеспечение (ПО) МетаФраз [10].

Лингвистическое программное обеспечение МетаФраз R10 (Metafraz Lingware R10) разработано в виде единого интегрированного многофункционального программного комплекса (Системы), состоящего из нескольких программных продуктов, предназначенных для решения отдельных функциональных задач в области компьютерной лингвистики.

В состав ПО МетаФраз R10, используемого для проведения экспериментов, входят следующие компоненты:

1) Библиотека словарей МетаФраз (MF Dictionary Lib R10) – основной ресурс Системы, содержащий комплекс декларативных (словарных) средств для задач фразеологического машинного перевода и семантической (смысловой) обработки текстов, а также набор грамматических таблиц для базовых лингвистических процедур.

2) Ядро лингвистического процессора и системы перевода (Kernel) – основной модуль Системы, включающий набор лингвистических программных библиотек, обеспечивающих выполнение всех лингвистических процедур Системы.

3) Лингвистический комплекс МетаФраз (MF Lingware Complex R10) – программный продукт Системы, входящий в состав автоматизированных систем МетаФраз, поддерживающих функционал создания и верификации словарей МетаФраз, включает модули создания частотных словарей по корпусу текстов, конвертации текстовых словарей в формат словарей МетаФраз и модуль Системы перевода МетаФраз.

4) Система семантической обработки текстов МетаФраз (MF Text Analyst R10) – программный продукт Системы, входящий в состав автоматизированных систем МетаФраз, поддерживающих функционал семантической (смысловой) обработки неструктурированных текстов на естественном языке, извлечения сущностей и установления связей, рубрикацию и кластеризацию документов, морфологический и семантических поиск и подбор документов, их автоматическое реферирование и перевод.

5) Электронная библиотека документов МетаФраз (база данных) – ресурс, входящий в состав Системы семантической обработки текстов МетаФраз, предназначенный для загрузки и хранения в БД документов (текстовых файлов) и результатов их лингвистической обработки. Электронная библиотека документов реализована с использованием СУБД MS SQL Server.

ПО МетаФраз обладает всеми необходимыми программными процедурами лингвистической обработки неструктурированных текстов, необходимых для решения данной задачи, и позволяет адаптировать декларативные средства для настройки на заданную предметную область путем быстрого автоматизированного создания словарей по корпусу текстов.

3.3 Автоматизированная настройка декларативных (словарных) средств на заданную предметную область

Для решения задач автоматической обработки информации в заданной предметной области необходимо провести работу по составлению семантических декларативных средств, в которых представляется понятийный состав предметной области и фиксируются смысловые отношения между понятиями.

Общая технологическая схема составления концептуального словаря представляется в следующем виде.

Предварительно составленный корпус текстов подвергается обработке процедурой семантико-синтаксического и концептуального анализа текстов, в результате чего из текстов выделяются отдельные слова и словосочетания различной длины. После этого по массиву выделенных из текстов слов и словосочетаний составляется частотный словарь.

Полученный словарь обрабатывается процедурой орфографического и синтаксического контроля, в результате чего из этого словаря исключаются некорректные слова и словосочетания. Частотная часть словаря подвергается лингвистической обработке, в результате которой из словаря исключается малоинформативная и некорректная лексика.

Далее выполняется автоматическое приведение наименований понятий к их канонической форме и формируется частотный словарь наименований понятий. И, наконец, на завершающем этапе выполняется семантико-статистический анализ частотного словаря на основе статистических данных о количественном и качественном составе этого словаря. С этой целью автоматически формируется характеристическая таблица частотного словаря. Для этого частотный словарь предварительно упорядочиваются по убыванию частот встречаемости слов в текстах и для каждой

частоты вычисляются такие параметры как его кратность, накопленная частота, накопленная кратность и относительная накопленная частота. Эти параметры позволяют выявить частотный понятийный состав предметной области и соотнести его с параметром покрытием этой частотой текстов предметной области.

Автоматизация составления словарей позволяет в короткие сроки и с минимальными трудозатратами создать для заданной предметной области систему взаимосвязанных наименований понятий, основанную на корпусе реальных текстов в предметной области.

3.4 Создание классификаторов в предметной области

Для создания классификаторов в предметной области были реализованы следующие этапы обработки корпуса текстов собранных публикаций:

1) разработаны методы, алгоритмы и ПО для выявления предикативных (глагольных) словосочетаний, характеризующих направление деятельности органов государственной власти и оценки их качества по корпусу текстов публикаций СМИ;

2) разработаны методы, алгоритмы и ПО для выявления оценочных суждений о деятельности органов государственной власти и оценки их качества (по корпусу текстов сообщений социальных сетей);

3) проанализирован корпус текстов СМИ, относящихся к тематике деятельности органов власти ХМАО, автоматизированным способом выделено свыше 2000 объектов, связанных со структурой органов власти региона и их деятельностью (государственными функциями);

4) по корпусу текстов СМИ выявлено более 6000 предикативных (глагольных) словосочетаний, характеризующих направление деятельности органов государственной власти и оценки их качества;

5) определены типы выделенных объектов и проведена их классификация, определяющая их отношение к структуре органов власти:

– губернаторская структура (губернатор лично, пресс-служба, аппарат, аффилированные лица губернатора);

– исполнительная власть (правительство, министерства, госпредприятия);

– законодательная власть;

– муниципальная (городские и районные власти, коммунальные службы, муниципальные предприятия) –

и их структурным подразделениям и выполняемым государственным функциям;

6) определены типы сообщений СМИ и публикаций в социальных сетях с точки зрения их отношения к объектам мониторинга:

– информирующие сообщения (преимущественно СМИ) в отношении конкретных информационных поводов (фактов, событий, персон);

– сообщения, демонстрирующие осведомленность населения по конкретным информационным поводам;

– отношение населения (позитивное, негативное, нейтральное) к конкретным информационным поводам.

На основе результатов этой обработки созданы классификационные словари объектов и отношений в заданной предметной области.

3.5 Обработка текстов средствами МетаФраз

После настройки декларативных средств лингвистического программного обеспечения на заданную предметную область мы провели автоматическую обработку имеющегося массива тестовых данных.

Для каждого документа было автоматически сформировано формализованное представление документа, включающее:

1. Список ключевых выражений со следующим набором данных по каждому из них:

а) ключевое выражение;

б) нормализованное ключевое выражение (пословно в канонической форме);

в) хеш выражения;

г) вес выражения в тексте с учетом предметной области;

д) группа (группы) по классификатору;

е) адреса в исходном тексте (координаты всех вхождений в тексте).

2. Общий автоматический реферат (автоконспект) заданного объема по документу.

3. Список объектов со следующим набором данных по каждому из них:

а) наименование объекта;

б) нормализованное наименование объекта (пословно в канонической форме);

в) хеш выражения;

г) вес выражения в тексте с учетом предметной области;

д) класс объекта;

е) адреса в исходном тексте (координаты всех вхождений в тексте).

4. Список объектов с предикатами со следующим набором данных:

а) объект;

б) предикат;

в) класс предиката (для типизации отношений к объектам мониторинга).

5. Список установленных связей между объектами:

а) объект 1;

б) предикат-связь;

в) объект 2.

Также было проведено последовательное отождествление документов по степени смысловой близости для кластеризации документов – группировка текстов публикаций по темам (информационным поводам).

Для отождествления смысловой близости документов в нашей задаче применялись не полные тексты документов, а их автоконспекты.

Автоконспект документа, представляющий собой автоматический реферат по наиболее значимым для данного текста и всей предметной области ключевым выражениям, позволяет провести группировку документов по их основному (наиболее значимому) информационному поводу. этот процесс также облегчается фиксированным (изначально заданным в настройках лингвистического процессора) объемом автоконспекта.

3.6 Оценка применимости полученных показателей

Представленное формализованное описание документов и методы его построения, а также методы кластеризации документов по информационным поводам позволили получить следующие данные, пригодные для проведения количественного анализа по текстам публикаций:

1) все сообщения (публикации СМИ, сообщения пользователей «ВКонтакте») сгруппированы по информационным поводам – что дает количественно измеряемый параметр частоты встречаемости информационного повода и его доли значимости в общем объеме публикаций;

2) все сообщения классифицированы по нескольким основаниям:

а) тип информационного сообщения (публикации СМИ, высказывания пользователей соцсетей);

б) положение основного объекта мониторинга, определяющего информационный повод, в иерархической структуре органов власти и их подведомственных организаций;

в) оценка основного объекта мониторинга (позитивная, негативная, нейтральная) –

что дает возможность проводить многомерный анализ сообщений по интенсивности воздействия СМИ, откликам пользователей сети, оценкам деятельности власти в разрезе иерархии органов власти и функциональной принадлежности предприятий и государственных услуг и т.п.

Расширение возможностей классификации и группировки сообщений с привлечением экспертов позволит существенно увеличить количество предоставляемых показателей, позволяющих проводить количественный статистический анализ данных.

4 Заключение

В настоящей статье авторы рассмотрели возможности создания формализованного представления содержания информационных сообщений для получения показателей количественной оценки деятельности органов власти по материалам электронных средств массовых коммуникаций.

Авторы описали структуру и методы создания формализованного описания документа для этой задачи, а также методы создания декларативных средств для получения наиболее адекватных результатов анализа в заданной предметной области.

Проведенные эксперименты на реальных данных показали жизнеспособность данного подхода, на основе которого можно создавать действующие системы мониторинга и семантического анализа информационных сообщений.

В то же время необходимо более серьезно прорабатывать методы решения поставленных задач с привлечением экспертов-аналитиков, специализирующихся в данной предметной области.

Авторы данной статьи продолжают работы по этой проблеме.

Литература

[1] Старовойтов А.В., Пошатаев О.Н., Прохоров С.Н., Хорошилов А.А. Методы автоматизированного составления и ведения словарей // Сб. Информатизация и связь / Центр информационных технологий и систем органов исполнительной власти. – 2013. – № 3. – С. 91–97.

[2] Богданов Ю.М., Пошатаев О.Н., Хорошилов А.А. Принципы создания высокопроизводительных систем обработки и анализа текстовой информации // Сб. Информатизация и связь / Центр информационных технологий и систем органов исполнительной власти. – 2013. – № 3. – С. 74–81.

[3] Пошатаев О.Н., Хорошилов А.А. Методы анализа текстов в технологиях «Big Data» // сб. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», XV Всероссийская научная конференция RCDL 2013, Ярославль, Россия, 14–17 октября. – С. 30–38.

[4] Белоногов Г.Г., Гиляревский Р.С., Селедков С.Н., Хорошилов А.А. О путях повышения качества поиска текстовой информации в системе Интернет // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский институт научной и технической информации РАН. – 2012. – № 8. – С. 15–22.

[5] Белоногов Г.Г., Гиляревский Р.С., Хорошилов А.А. Проблемы автоматической смысловой обработки текстовой информации // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский институт научной и технической информации РАН. – 2012. – № 11. – С. 24–28.

[6] Белоногов Г.Г., Гиляревский Р.С., Хорошилов А.А., Хорошилов-мл. А.А. Автоматическое распознавание смысловой близости документов // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский институт научной и технической информации РАН. – 2011. – № 7. – С. 15–22.

[7] Белоногов Г.Г., Гиляревский Р.С., Хорошилов Ал-др А., Хорошилов Ал-ей А. Развитие систем автоматической обработки текстовой информации // Нейрокомпьютеры: разработка, применение. – 2010. – № 8. – С. 4–13.

[8] Белоногов Г.Г., Хорошилов Ал-др А., Хорошилов Ал-ей А. Единицы языка и речи в системах автоматической обработки текстовой // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский институт научной и технической информации РАН. – 2005. – № 11. – С. 21–29.

[9] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. Теория и практика построения систем автоматической обработки текстовой информации. – Москва: Информационно-издательское агентство «Русский мир», 2004. – 247 с.

[10] Сайт МетаФраз. http://www.metafraz.ru

Methods for Automatic Construction of a Formalized Representation of the Contents of Electronic Mass Communication Materials to Solve the Problem of Monitoring and Assessment of Authorities

Yury V. Nikitin, Alexander A. Khoroshilov, Alexei A. Khoroshilov

This paper addresses to the possibility of generating a formalized representation of the information publications in the Internet to derive a quantitative evaluation of the authorities based on the content of such publications. It also covers methods of constructing a representation of messages and methods of adaptation of automated semantic processing tools for obtaining the most appropriate analysis results in a given domain.