# Обнаружение информационных трендов на русском языке в сети интернет

А. В. Экало[1], С. А. Беляев[2]
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»
им. В.И. Ульянова (Ленина)
[1]ekalo@nicetu.spb.ru, [2]bserge@bk.ru

***Аннотация*. Авторы рассматривают эволюцию подхода к анализу информационных трендов в сети интернет на русском языке. Современный уровень технологий позволяет решать сложные и разнообразные задачи по анализу социальных сетей, новостных ресурсов в сети интернет и страниц отдельных авторов. В зависимости от выбранного подхода могут быть получены различные результаты – от статистических до результатов семантического поиска. Авторы предлагают модификацию существующих решений, описывают соответствующую математическую модель и предлагают алгоритм работы системы анализа информационных трендов в сети интернет.**

***Ключевые слова: семантический анализ текста на русском языке; системы сбора и обработки; децентрализованные хранилища данных***

## I. Эволюция подходов к анализу текста

Современные технологии существенно изменили нашу жизнь и вот мы уже не только используем технику как исполнительные механизмы, но и ожидаем интеллектуальный отклик с учётом нечеткости или неточности нашего запроса. Изменился и механизм взаимодействия, раньше пользователю необходимо было нажимать кнопки, набирать текст, а теперь мы можем с помощью голоса обратиться к компьютеру, телефону или телевизору и получить адекватную обратную связь [1, 2].

Развитие технологий взаимодействия с электронными устройствами обусловлено существенным прорывом в анализе естественного языка. Современные вопросно-ответные системы выполняют не только морфологический, лексический и синтаксический анализ, но и обеспечивают элементы семантического анализа [3], претендуя на возможность понимания полученного запроса.

Первые поисковые системы предлагали достаточно примитивную обработку текста, как максимум могли выполнить морфологический анализ и использовать специализированные правила формирования запроса. Со временем они стали более интеллектуальными и стали предлагать варианты того, что, скорее всего, ищет пользователь. Известные поисковые системы, такие как Google, Yandex и др., предлагают, в том числе, обработку запросов на естественном языке. Создаваемые ими инструменты на некоторые простые вопросы могут давать адекватные ответы, имитируя взаимодействие с настоящим помощником, а не обычной поисковой системой.

В целом русский язык сложнее для автоматической обработки чем английский язык. В настоящее время наблюдается активный рост числа исследователей, предлагающих свои решения по обработке вопросов на русском языке. В том числе проводятся соревнования (https://sdsj.ru/), на которых в 2017 году участникам необходимо было создать вопросно-ответную систему.

Сеть интернет предлагает широкие возможности по распространению информации с использованием новостных сайтов, социальных сетей, блогов и других инструментов обмена сообщениями. Все они могут носить как легальный, так и не легальный характер с учётом требований законодательства той или иной страны [4]. В связи с этим у контролирующих органов должны быть инструменты, позволяющие в автоматизированном, а лучше – автоматическом, режиме анализировать тексты и предоставлять исходные данные для принятия решений. В идеальном варианте средства анализа должны позволять обрабатывать не только сам текст, но и учитывать его эмоциональную окраску. В настоящее время разрабатываются решения, позволяющие определять, например, сарказм, аллегорический текст и т.п., но пока полностью автоматически без привлечения человека они не работают. Важным фактором, определяющим необходимость средств автоматического анализа текста, является быстро растущие объёмы новостной информации, большое количество индивидуальных авторов. Всё больше новых пользователей подключается к сети интернет, всё большее влияние эта сеть оказывает на массовое сознание. Не секрет, что зачастую пользователи перепечатывают чужие информационные сообщения или создают свои на основе существующих новостей. Перед контролирующими органами встаёт задача выявления источников информационных трендов и определения степени их влияния на пользователей сети интернет.

## II. Существующие решения

Разные авторы предлагают различные подходы, которые в той или иной мере решают задачу создания систем, обеспечивающих анализ текстов на естественном языке. Научные изыскания в большинстве своём ведутся

по узким направлениям особенностей русского языка, но предлагаются и всеобъемлющие подходы.

Использование онтологических моделей позволяет описать все понятия предметной области и связи между ними, сформировать структуру, которая будет представлять собой семантическую сеть какой-то предметной области. Данный подход является очень близким к восприятию пользователем с семантической точки зрения [5]. Вопрос при этом формируется с учётом семантической сети и данных, в ней хранящихся. К сожалению, данный подход имеет ограниченное применение, это связано со сложностью формирования качественной онтологической модели. Современные подходы по автоматическому формированию моделей позволяют описывать достаточно объёмные предметные области, но при этом пока уступают в качестве моделям, созданным квалифицированным оператором.

Анализ текста может осуществляться с использованием подходов, обеспечивающих разметку семантических ролей, но для русского языка, в отличие от английского, соответствующий семантический корпус в полном объеме не сформирован [6] или отсутствует в открытом доступе. Использование объёмного семантического корпуса существенно повысит качество вопросно-ответных систем. В настоящее время сосуществуют две основные парадигмы: вопросно-ответная система на основе извлечения информации (IR-based QA) и вопросно-ответная система, основанная на знаниях (Knowledge-based QA). Обе они применимы при анализе информационных трендов, соответственно, гибридная система, учитывающая ограниченность корпуса на русском языке, предложенная в [6], может использоваться в качестве инструмента семантического анализа текста.

В настоящее время продолжают использоваться подходы по анализу текстов из сети интернет с привлечением экспертов [7]. При этом оценивается не только материал, запрашиваемый пользователем, но и адекватность оценки эксперта с выставлением соответствующего рейтинга. Данный подход позволяет максимально точно оценивать текст и позволяет учитывать все особенности языка, которые могут быть пропущены при автоматической обработке, но у него есть и существенный недостаток – его высокая трудоёмкость.

Для автоматического выделения ключевых понятий и аннотирования текста существует множество библиотек для английского (Google Natural Language API, NLTK, Stanford NER, OpenNLP, ParallelDots, Spacy, Aylien, TextRazor) и русского (Abbyy Infoextractor, Natasha, DaData, Pullenti, Promt, RCO, Ahunter) [8] языков, они используют настраиваемые правила для каждого понятия. Для русского языка в настоящее время активно развивается библиотека Natasha (http://natasha.readthedocs.io), она работает с использованием грамматик и словарей для парсера Yargy. Применение данных библиотек позволяет обнаруживать упоминание тех или иных понятий в анализируемом тексте, при незначительной доработке можно выяснить контекст упоминания искомых понятий и использовать эти данные для обнаружения новых информационных трендов.

Публикации в сети интернет отличаются тем, что их пишут не только профессиональные авторы, но и авторы, которые не всегда следуют канонам языка. С учётом этих особенностей возможны ситуации, когда в слове появятся ошибки и опечатки, и оно будет не корректно сравниваться с искомым. В связи с этим для анализа информационных трендов может использоваться нечеткий поиск [9], который позволяет минимизировать влияние грамматических ошибок на результаты поиска. В [9] предлагается подход с использованием свободно распространяемого программного обеспечения для решения задачи нечеткого поиска в сети интернет. В основе предложенного решения находится Apache Lucene – высокопроизводительная библиотека для полнотекстового поиска. По своему функционалу она близка к возможностям поисковых роботов.

Поисковые роботы обеспечивают сканирование и обработку большого количества источников в сети интернет. Результаты сканирования могут использоваться для решения различных задач, например, с использованием методов выделения ключевых понятий в тексте можно проанализировать количество упоминаний в сети интернет имен участников предвыборной гонки в зависимости от различных событий (https://habrahabr.ru/post/351040/). При этом с заданной периодичностью сохраняются результаты работы поисковых роботов. Изменение количества упоминаний в разные моменты времени говорит о влиянии анализируемых событий. Данный подход является статистическим, информация о событиях вносится оператором. Максимум, что можно получить как результат, – степень влияния события. Если оператор не знает о событии, то он его не обнаружит в выборке.

Инструмент обнаружения новых информационных трендов в сети интернет должен как минимум обеспечивать решение следующих задач:

- сканирование и обработка большого количества источников в сети интернет, хранение истории результатов поиска для организации сравнения с результатами предыдущих поисков;
- автоматическое выделение ключевых понятий с использованием алгоритмов нечёткого поиска;
- применение алгоритмов семантического анализа текстов для выявления связей между понятиями и контекста их употребления;
- построение графов распространения информации в сети интернет.

В [9] описаны способы построения графов распространения информации, но в предложенном подходе отсутствует семантический анализ текста и не используются алгоритмы извлечения знаний. Следует отметить, что с учётом развития современных технологий появились возможности по децентрализованному хранению данных [10].

Децентрализованные приложения обеспечивают децентрализованное хранение информации и отсутствие центральной точки отказа. При этом существенно изменяется механизм доступа к данным и доступ даже на чтение может оказаться не бесплатным. В настоящее время основное применение децентрализованных приложений – обеспечение операций с криптовалютой, хранение данных и так называемые смарт-контракты [10], но это не означает, что с использованием того же механизма не могут храниться тексты или описания смарт-контрактов. Таким образом децентрализованные приложения могут оказаться новым нестандартным источником информации при анализе информации из сети интернет.

## III. Интеграционный подход

Возможны различные варианты последующей визуализации результатов анализа информационных трендов в сети интернет [8, 9], но при этом возникает потребность обеспечить максимальное качество достигнутых результатов. Для этого целесообразно модифицировать предложенную в [9] математическую модель процесса анализа информационных трендов: M = (A, G, C, L, R, T), где A = {a, dc} – список Интернет-адресов новостных сайтов, социальных сетей, блогов и децентрализованных хранилищ данных (dc – конфигурация и ключи для доступа к информации), G:A→H – функция получения HTML-кода с сайтов с учётом внутренних переходов между страницами по ссылкам, здесь H = {h} – множество полученных HTML-страниц. C:H→L – функция предварительного анализа и формирования множества L = {l = <a,d,v,n>}, внутреннего представления публикаций, здесь a – адрес публикации, d – дата её обнаружения поисковым роботом, v – текст, подготовленный для анализа (подготовленный в том числе для использования алгоритмов нечёткого поиска), n – семантическая сеть, описывающая текст. T = R(L) – множество обнаруженных информационных трендов, сформированных по элементам множества L. Каждый элемент множества T = $\{t_i\}$ представляет собой кортеж $t_i = <d_i,gr_i,h_i>$, где $d_i$ – дата и время появления информационного тренда, $gr_i$ – граф распространения информации в сети Интернет, $h_i$ – HTML-текст первоисточника информационного тренда. Граф распространения информации в узлах содержит ссылки на публикации, по дугам осуществляется переход к информации о сайтах, на которых публикация появилась позже.

Предложенный подход отличается от описанного в [9] возможностью получения информации из децентрализованных хранилищ данных, построением семантической сети анализируемого текста, сравнение текстов не на основе статистических алгоритмов и учитывает, что информация от поисковых роботов поступает регулярно и следует хранить не вычисленную дату появления информации, а дату её обнаружения поисковым роботом. При этом не требуется внесение изменений во внутреннее представление документа, в описание индекса документа, в перечень предложенного к использованию свободного программного обеспечения. Существенные изменения возникают в реализации интерфейса получения данных из Интернета, т.к. появились новые не предусмотренные источники данных (децентрализованные хранилища). В данном случае не предполагается, что описываемая система будет самостоятельно выступать в роли децентрализованного хранилища, она обращается к этим хранилищам и получает доступ к хранимым данным на чтение. Существенные изменения вносятся в алгоритмы предварительного анализа документа, т.к. предполагается, что будет выполнен семантический анализ текста, который позволит обеспечить более точную проверку графа распространения информации. Предлагается отказаться от статистического сравнения двух текстов, с учётом существенного развития алгоритмов семантического анализа текста предлагается использовать их для определения степени сходства двух текстов, что позволит принять решение о заимствовании данных. По сравнению с [9] существенно изменяется принцип сохранения информации в БД. Предлагается сохранять данные, полученные поисковым роботом с обозначением времени их получения, а не пытаться автоматически определять время появления информации, что позволит избежать ошибок, связанных с настройкой web-серверов, предоставляющих данные. В данном случае это позволит в том числе избежать попыток умышленно изменить время появления информации.

Общая последовательность работы предлагаемой системы для анализа информационных трендов в этом случае следующая:

Шаг 1. Перебор в цикле источников данных.

Шаг 1.1. Получение HTML-документа.

Шаг 1.2. Разбор HTML-документа.

Шаг 1.3. Формирование внутреннего представления документа

Шаг 1.3.1. Формирование семантической сети.

Шаг 1.3.2. Фиксация даты получения информации.

Шаг 1.4. Сравнение текстов на основании их внутреннего представления.

Шаг 1.5. Группировка документов по результатам сравнения.

Шаг 1.6. Формирование графов распространения информации с использованием даты обнаружения документов в сети Интернет.

Шаг 2. Получение запроса от пользователя.

Шаг 3. Отображение пользователю графа распространения информации.

## IV. Заключение

Предложенный подход является модификацией существующих решений и учитывает, как современные тенденции по публикации информации в сети интернет, так и новейшие методы анализа текстов на русском языке.

Логичным следующим шагом является проведение соответствующих экспериментов и сравнение результатов.

## Список литературы


[1] Автоматизированная система выявления нештатных ситуаций в процессах управления сложными техническими системами и поиска решений по их устранению / Беляев С.А., Романенко Д.А. // Тез.докл. 7-го Всеросс. совещ. по проблемам управления ВСПУ-2014, Москва, 16–19 июн 2014. / М.: Институт проблем управления им. В.А. Трапезникова РАН, 2014. С. 8612-8619.

[2] Нормализация текста в системе русскоязычного синтеза «текст-речь»: классификация и обработка нестандартных текстовых объектов / Черепанова О.Д. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 31 мая – 3 июня 2017 г.). Вып. 16 (23): В 2 т. Т. 1. М.: Изд-во РГГУ, 2017. С.42-51.

[3] Смирнов И.В., Шелманов А.О., Кузнецова Е.С., Храмоин И.В. Семантико-синтаксический анализ естественных языков. Ч. II: Метод семантико-синтаксическогоанализа текстов. М.: Изд-во ИПИ РАН, 2014. URL: http://www.aidt.ru/images/documents/2014-01/11_24.pdf (дата обращения: 15.03.2018).

[4] Подход к построению комплексной системы предупреждения преднамеренных информационных трендов на основе семантического анализа текстовых ресурсов в сети Интернет / Солодухин А.И., Романенко С.А., Беляев С.А., Медведева Я.И. // Актуальные проблемы психологической безопасности: Тез.докл. регионального совещания. СПб, 5 июн 2012. СПб: Изд-во «Свое Издательство», 2012. С. 79-86.

[5] Ontology and Integration of Formal and Lexical Semantics / Borschev V.B., Partee B.H. // Компьютерная лингвистика и интеллектуальные технологии: Тез.докл. Международной конференции «Диалог», Бекасово, 4–8 июн 2014. Вып. 13 (20). М.: Изд-во РГГУ, 2014. С. 114-127.

[6] Belyaev S.A., Kuleshov A.S., Kholod I.I. Solution of the Answer Formation Problem in the Question-Answering System in Russian (доклад) // Institute of Electrical and Electronics Engineers Inc. 2017 IEEE Russian Young Researchers in Electrical and Electronic Engineering Conference, ElConRus 2017, SPb, 1-3 Feb 2017, SPb: Saint Petersburg Electrotechnical University "LETI". PP. 360-365.

[7] Супруненко А.В. Способ экспертной оценки с использованием сети репутации для решения задачи классификации веб-контента // Искусственный интеллект и принятие решений, 2016. №3. С.72-76.

[8] Заболеева-Зотова А.В., Орлова Ю.А., Розалиев В.Л. Комплексный семантический анализ потока новостных текстов // Искусственный интеллект и принятие решений, 2015. №4. С.81-88.

[9] Беляев С.А., Васильев А.В., Кудряков С.А. Архитектура системы мониторинга информационных трендов на основе свободного программного обеспечения // Программные продукты и системы. 2016. Т.29 № 4. С. 85–88.

[10] Равал С. Децентрализованные приложения. Технология Blockchain в действии. СПб: Питер, 2017. 240 с.