Д.А. Девяткин, Р.Е. Суворов, И.В. Соченков

Архитектура поисково-аналитической системы и исследование информационного пространства, связанного с арктической зоной[1]

Аннотация. В работе рассмотрен вопрос создания методов и экспериментальных программных средств информационной поддержки принятия решений на основе анализа информационного пространства для заданной темы. Приводится постановка задачи создания таких программных средств, их архитектура и основной алгоритм работы, типы и структура источников информации, а также типы запросов, которые могут быть полезны эксперту. В качестве примера рассматривается информационный фон вокруг арктической зоны. Приведена методика и результаты первичного анализа информационного пространства, включающие описание актуальных сюжетов и ассоциативных правил, описывающих значимые связи между сущностями.

Ключевые слова: информационно-поисковая система, мониторинг информационного пространства, мониторинг событий, извлечение информации, извлечение отношений, базы знаний, поддержка принятия решений.

Введение

Информация, извлекаемая из сообщений СМИ и социальных сетей, часто используется для принятия решений. Существуют специализированные системы, решающие задачи из достаточно узкого спектра, например отдельные задачи в рамках поддержки спасательных операций или борьбы с криминалом в определенном регионе [1]. В виду специализации и особенностей существующего подхода, основанного на анализе частот сообщений по каждой тематике и их распределения во временных интервалах, универсальные системы, допускающие сложные запросы, отсутствуют.

В настоящей работе исследуются интеллектуальные методы обнаружения и выделения описания событий, относящихся к заданной теме, из информационных потоков интернета на примере арктической зоны. В работе предлагается подход к интеграции методов интеллектуального анализа данных и текстов для решения задачи информационной поддержки принятия решений на основе анализа информационного пространства по заданной теме. Новизна подхода заключается в решаемой задаче и интегральном подходе. Особенностью задачи является предоставление эксперту возможности задавать широкий спектр вопросов к базе знаний, например “Какие события происходили в такое-то время и/или в таком-то месте?”, “Кто участвовал в таких-то событиях?”, “В каких событиях ещё участвовали стороны, участвовавшие в заданном событии?” и т.п. При этом база знаний интегрирует как структурированные данные, полученные посредством извлечения именованных сущностей, отношений между ними, так и временную и географическую привязку (посредством геотеггинга).

В качестве источников этих информационных потоков выступают сайты информационных агентств, блоги, социальные сети, форумы и другие интернет-ресурсы. Из этих источников был сформирован тестовый набор данных, а

[1]Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 15-29-06031 офи-м, № 15-29-06082 офи-м.

также проведен их первичный анализ. Выполнена экспериментальная оценка этого подхода на тестовом наборе данных.

Рассмотрена также архитектура программных средств информационной поддержки принятия решений на основе анализа информационного пространства по заданной теме.

Связанные работы

Данная работа относится к области методов для специализированных поисково-аналитических машин, предназначенных для непрерывного мониторинга информационного пространства и его анализа, включающего в себя получение информации о происходящих событиях, действующих сторонах (странах, организациях, физических лицах), высказываниях, договорённостях, аффилиациях и т.п.

Задача автоматического обнаружения различных событий по интересующей теме на основе анализа информационного потока СМИ и социальных медиа исследуется в работе [2]. В этой работе предложен вычислительно эффективный метод обнаружения первого упоминания целевого события в СМИ.

В статье [3] описывается система TEDAS, направленная на выявление и анализ различных резонансных событий по сообщениям в социальной сети Twitter. В этой системе для настройки сфокусированного сбора данных применяется обучение с подкреплением.

Во многих исследованиях изучаются проблемы создания систем поддержки принятий решений при ликвидации последствий чрезвычайных ситуаций, основанных на мониторинге социальных медиа. Так, например, в исследовании [4] представлена система мониторинга информации о чрезвычайных происшествиях, в состав которой входит сфокусированный краулер (web-crawler). Для организации сфокусированного обхода и загрузки данных из источников информации авторами создана онтология, посвященная различным чрезвычайным происшествиям. В работе [5] рассматривается платформа Tweedr, которая используется для информационной поддержки спасателей. В этой системе из сообщений Twitter извлекаются именованные сущности, такие как пострадавшие объекты, типы чрезвычайных ситуаций и др. Был также реализован классификатор, который предоставляет спасателям только те сообщения, в которых имеются указания на ущерб инфраструктуре или на наличие пострадавших. Для извлечения информации из текстов сообщений применяются условные случайные поля(Conditional Random Fields) [6].

При создании систем поддержки принятия решений часто применяются онтологии предметной области [7, 8]. Однако формирование онтологии по широкой теме, например, такой как ситуация в Арктической зоне, является нетривиальной задачей. В отличие от предметной области с очерченной терминологией и соответствующей лексикой, подобная тематика может включать в себя информацию, относящуюся к разным предметным областям в рамках темы. Обширность подобных тем требует создания новых методов работы с терминологией и лексикой для обнаружения целевой информации из текстов. Составляемые вручную словари, тезаурусы и онтологии могут служить лишь отправной точкой для решения такой задачи. При этом необходимо учитывать их изначальную неполноту и тенденцию к быстрой потере актуальности из-за возникновения и развития новых «подтем» в рамках общей темы.

Рассмотрим базовые функциональные элементы, присущие системам поддержки принятия решений на основе мониторинга СМИ. К этим элементам относятся подсистемы сбора данных по интересующей теме, извлечения именованных сущностей и полнотекстового поиска.

Одним из широко применяющихся подходов к сфокусированному сбору данных является алгоритма shark search [9]. Этот алгоритм принимает в качестве входных параметров поисковый запрос и набор веб-страниц, являющихся начальными точками сбора информации (*seeds*). Эти страницы анализируются на предмет релевантности запросу, также из них извлекаются ссылки на другие страницы. Для каждой ссылки также вычисляется мера её релевантности запросу. Затем извлеченные ссылки добавляются в очередь с приоритетами, причем приоритет каждой ссылки вычисляется на основе релевантности её родительской страницы и релевантности её контекста. В случае если релевантность ссылки снижается ниже определенного порога – она не добавляется в очередь, т.е. сбор данных в этом направлении останавливается. На следующем шаге из очереди извлекается очередная ссылка и по ней за-

гружается страница. Алгоритм продолжает свою работу пока очередь не опустеет. В исследовании [10] представлены способы улучшения производительности этого алгоритма.

Другим подходом к организации сфокусированного сбора информации является интеллектуальный краулинг [11]. Суть этого подхода состоит в поиске в исходном запросе концептов заранее сформированной онтологии анализируемой предметной области и расширении этого запроса близкими концептами. Авторами вводится метрика семантической близости, которая вычисляется на основе сходства концептов в онтологии. На основе этой метрики вычисляется релевантность загружаемых страниц запросу.

Еще один подход к сфокусированному сбору информации состоит в применении методов машинного обучения, настроенных на больших размеченных наборах веб-страниц, для предсказания корректных переходов между страницами в процессе краулинга. В работе [12] в качестве такого метода машинного обучения предлагается использовать скрытые марковские модели.

Зачастую необходимо рассматривать не страницы целиком, а отдельные их фрагменты (например, статьи и комментарии к ним). Существуют работы (например, [13]), применяющие автоматически генерируемые по примерам правила для извлечения значимых фрагментов HTML-страниц.

Для извлечения именованных сущностей из текстов часто используются методы обучения с учителем. В исследовании [14] представлен метод, основанный на обучении с учителем. Такой подход позволяет добиться высокой точности извлечения сущностей, однако требует разметки объемных обучающих наборов данных. Этот недостаток становится значимым, когда требуется организовать извлечение именованных сущностей из мультиязычного набора документов.

Подход с частичным обучением с учителем, не требующий применения размеченных корпусов текстов, представлен в статье [15]. Этот подход состоит в следующем. Во-первых, для всех слов, взятых из Википедии [16], при помощи многослойной нейронной сети строятся их расширения, которые неявно кодируют семантические и синтаксические свойства этих слов. Во-вторых, из каждой статьи в Википедии извлекаются контексты ссылок на другие статьи. Если эти статьи определяются во Freebase [17] как некоторые именованные сущности, то их контексты помечаются как положительные обучающие примеры для классификатора. Далее обученный классификатор может использоваться для извлечения сущностей.

Наиболее распространенным подходом к реализации полнотекстового поиска является использование инвертированных поисковых индексов [18]. В работе [19] полнотекстовые инвертированные индексы были усовершенствованы для организации вычислительно эффективного поиска, учитывающего наличие синтаксических и семантических связей между элементами текста.

Парадигма открытого извлечения информации может быть полезна при построении обобщённой системы информационной поддержки принятия решений. Методы и системы открытого извлечения информации направлены на извлечение бинарных отношений между сущностями из массивных текстовых коллекций (в виде набора триплетов <сущность, отношение, сущность>). При этом множество отношений заранее не оговаривается, а считается, что каждый глагол может выступать в качестве идентификатора отношения. Так, в исследовании [20] рассматривается система извлечения отношений, использующая статистически отобранные синтаксические шаблоны для извлечения триплетов. В работе [21] рассматривается подход к применению результатов лингвистического анализа текста для извлечения триплетов, основанный на обучении классификатора, управляющего конечным автоматом, осуществляющим обход синтаксического дерева. Актуальность данной проблемы подтверждает регулярное проведение публичных соревнований по этой тематике, таких как соревнование по автоматическому построению базы знаний KBP [22]. Существенной проблемой в рамках открытого извлечения информации является разрешение конфликтов и объединение синонимичных связей [23]. К тому же, полнота и точность извлечения отношений далеки от идеальных и находятся на уровне 0,6 – 0,7.

Таким образом, не смотря на множество работ, посвящённых отдельным этапам обработки информации в информационно-аналитических системах, можно сделать заключение об актуальности данной работы в виду отсутствия комплексных методов и интегрированных ин-

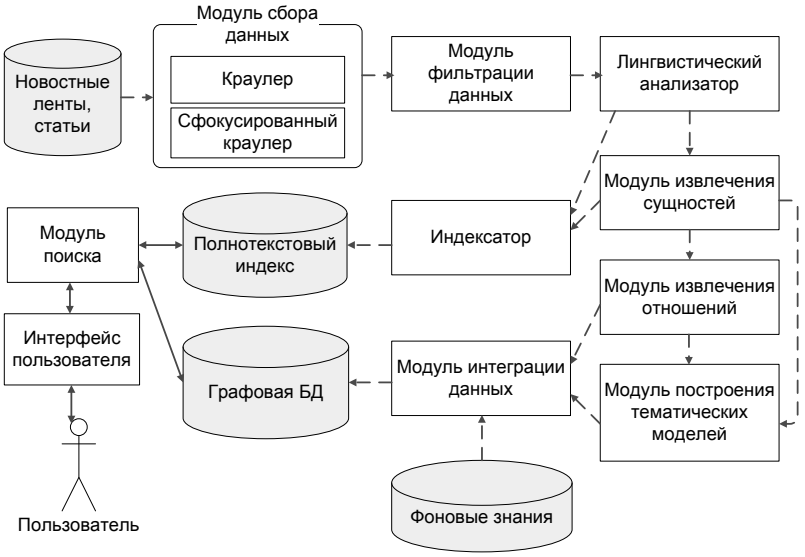

Рис. 1. Архитектура системы информационной поддержки принятия решений

формационно-аналитических систем, предоставляющих функции информационной поддержки принятия решений в рамках широкой темы (без ориентации на конкретные типы событий, как в системах, направленных на помощь в экстренных ситуациях).

Архитектура программных средств

Разрабатываемые программные средства информационной поддержки принятия решений на основе анализа информационного пространства состоят из следующих основных структурных компонентов (Рис. 1).

Модуль сбора данных выполняет функции загрузки информационных ресурсов по заданной теме. В качестве таких ресурсов предполагается использовать RSS ленты новостных агентств, сайты СМИ, научных и научно-популярных журналов. Для загрузки данных из ресурсов, посвященных исключительно заданной теме, будет использоваться универсальный веб-краулер, основанный на платформе Scrapy [24]. Сбор данных из ресурсов, посвященных широкому спектру различных тем будет производиться с помощью сфокусированного краулера.

Модуль фильтрации данных предназначен для отделения информативных текстов, тематически релевантных заданной теме, от служебной или рекламной информации. Предлагается реализовать грубый тематический классификатор, допускающий ошибки второго рода (ложные положительные), но при этом практически не допускающий ложноотрицательных решений (пропусков релевантной информации). В качестве такого классификатора может служить поиск терминологии, собственных наименований и ключевой лексики по заданной тематике.

Лингвистический анализатор Exactus [25] используется для выполнения полного семантико-синтаксического разбора текста. Результатом анализа является формальное представление текста в виде неоднородной семантической сети.

Модуль извлечения именованных сущностей выполняет выделение из текстов объектов различных типов, таких как организации, географические локации, страны, военные объекты, природные ресурсы, персоны и др. Для решения этой задачи предполагается разработать комбинированный подход, в котором применяется метод с частичным обучением с учителем [15] и метод, использующий дополнительные лингвистические ресурсы (тезаурус) для выделения таких сущностей как организации, географические локации, персоналии, военные объекты, должности и др.

Модуль извлечения отношений решает задачу выявления различных связей между сущностями, в том числе действия (связи, определяемые глаголами), аффилиации, ассоциации. Предполагается, что модуль будет использовать комбинированный алгоритм, использующий результаты синтаксического и семантического анализа (элементы предикатных структур) для первичного выделения отношений и отбора отношений с использованием статистики [26].

Извлечённые отношения могут использоваться как для построения базы фактов, так и для полуавтоматического построения классификатора ассоциативных правил, который в дальнейшем может использоваться для определения потенциально интересных экспертам ситуаций.

Индексатор Exactus [19] выполняет обработку и сохранение текстов в поисковом инвертированном индексе.

В модуле построения тематических моделей выполняется выделение основных тематик, содержащихся в анализируемых текстах. Для этого будет использоваться метод латентного размещения Дирихле (LDA) [27], а также авторский метод, более подходящий для анализа больших коллекций [28]. При этом будет выполняться несколько итераций кластеризации с различной гранулярностью для повышения полноты и точности извлечения сюжетов из информационного потока.

Модуль интеграции данных отвечает за объединение результатов извлечения именованных сущностей и отношений из различных источников в единую базу знаний, представляющую собой гетерогенную сеть. База содержит только те факты, которые можно непосредственно извлечь из текстов, либо с помощью сопоставления сущностей записям имеющихся тезаурусов (“*Фоновые знания*” на Рис. 1). Основные сущности, их атрибуты и допустимые типы связей между ними изображены на Рис. 2. Не все связи помечены метками ввиду очевидности их смысла, либо в виду зависимости смысла от ситуации. Пунктирные линии, ведущие от группы сущностей “*Действие*” обозначают, что эти сущности являются элементами множества отношений различной арности (не менее двух). При этом различных действий может быть множество, каждое действие соответствует некоторой группе близких по смыслу глаголов, и обладает своим набором характерных связей. Под “*Ресурсами*” понимаются объекты, которые могут выступать предметом конфликта интересов или просто предметом действия (нефть, газ, редкие животные и т.п.). Под “*Мероприятиями*” понимаются встречи, конференции, подписание договоров, приобретение, слияние, бурение, разведка месторождений и т.п. Всем связям, сущностям “*Событие*” и “*Действие*” ставится в соответствие момент времени, в которые, как предполагается, соответствующие факты имели место быть. Сущностям “*Локация*” сопоставляются географические координаты для отображения на карте. Географическая привязка осуществляется с помощью OpenStreetMap [29]. Все сущности и связи имеют ссылки на документы в полнотекстовом индексе, в которых они были обнаружены.

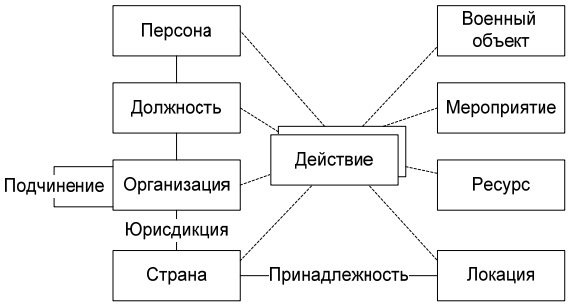

Рис. 2. Основные сущности и связи между ними

База фактов хранится в графовой базе данных Titan [30], совмещённой с хранилищем Cassandra [31]. Преимущество данной базы данных в особой структуре внутренних индексных структур. В отличие от реляционных СУБД, в которых принято для каждого запроса строить отдельный индекс, покрывающий все записи таблицы, в Titan существует несколько уровней индексов: глобальные индексы для поиска вершин, локальные индексы для поиска инцидентных рёбер и соседних вершин (хранятся для каждой вершины по отдельности). Такая структура индексов позволяет вычислительно эффективно осуществлять глубокие обходы, содержащие множество переходов. В РСУБД это привело бы к так называемому join-взрыву и отказу в обслуживании [32].

Модуль поиска осуществляет преобразование описания поисковых критериев, задаваемых пользователем посредством графического интерфейса, в набор конкретных запросов к базе фактов и полнотекстовому индексу, а также осуществляет слияние результатов поиска из разных источников. Для поиска в базе фактов применяется язык запросов Gremlin [33].

Все компоненты либо поддерживают параллельное исполнение, либо были специально спроектированы для запуска в распределённой среде (например, Titan и Cassandra), поэтому в целом вышеописанная архитектура позволяет организовать обработку и поиск практически неограниченных объёмов информации.

Общий алгоритм работы программных средств

Рассмотрим процесс пополнения информационных баз (полнотекстового индекса и графовой БД) в представленной системе (Рис. 3). Текстовые данные, посвященные исследуемой тематике и собранные из новостных лент, сайтов СМИ и журналов разбираются лингвистическим анализатором, для всех элементов текста определяются их морфологические, синтаксические и семантические характеристики. Затем выполняется извлечение именованных сущностей, определяются элементы текста, соответствующие наименованиям географических локаций, стран, военных объектов, организаций, природных ресурсов, должностей и персон. Далее производится индексация этих текстовых данных и извлечение отношений, выражающих связи между сущностями. Генерируются тематические модели, характеризующие набор актуальных сюжетов в анализируемой предметной области. Затем элементы полученной базы знаний сохраняются в графовой БД.

Общий алгоритм работы является достаточно типичным для такого рода систем, при этом основным его отличием является интеграция методов интеллектуального краулинга, методов глубокого анализа текстовой информации, статистических алгоритмов генерации правил, а также современных систем хранения больших объёмов данных.

Первичное исследование связанного с арктической зоной информационного пространства

Основная цель первичного исследования – получить представление об основных элементах информационного пространства, связанного с крайним севером и арктической зоной, сюжетах, типичной лексике, действующих лицах и пр. Другой целью исследования является проверка предложенной архитектуры и алгоритмов. Методика проведения первичного исследования частично повторяет описанный общий алгоритм работы системы. При этом для ускорения получения результатов некоторые этапы обработки упрощены (например, не выполняется синтаксический и семантический анализ).

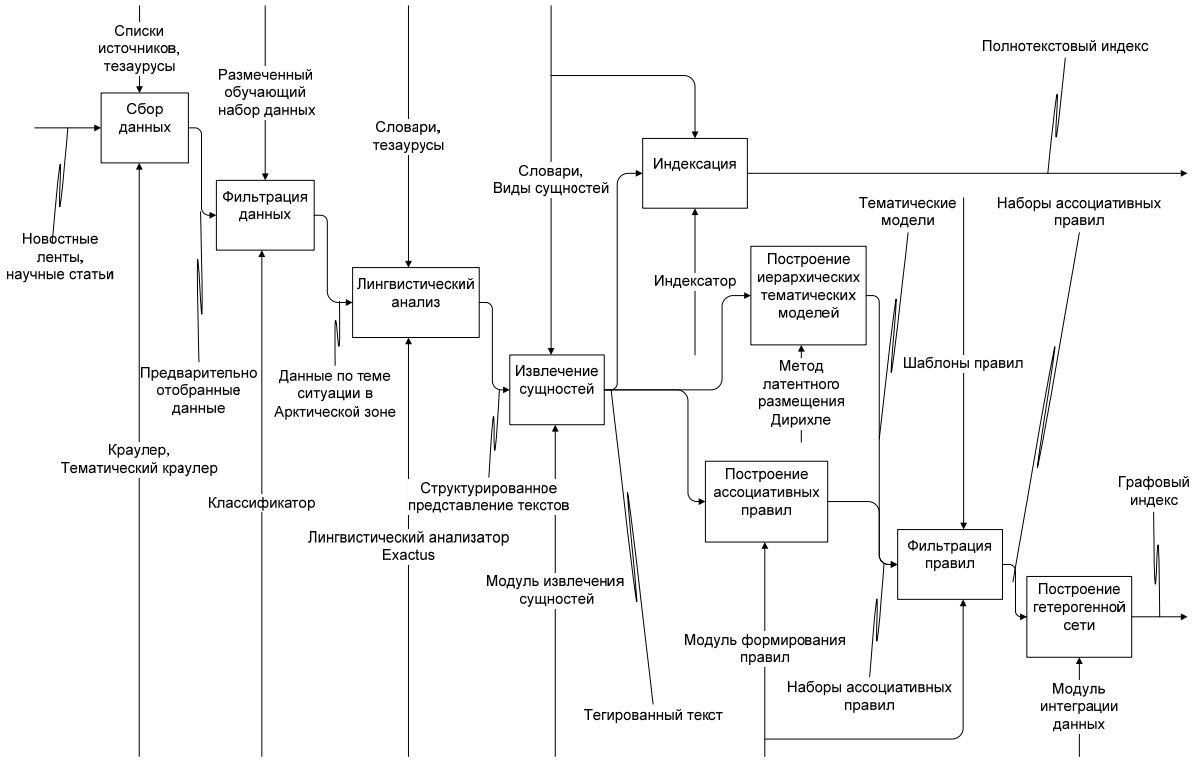

Рис. 3. Диаграмма процесса наполнения информационной базы системы

В рамках первичного исследования было собрано более 18 тысяч документов на русском (14 тысяч) и английском (4 тысячи) языках. В качестве источников в основном выступали специализированные новостные, аналитические или научные информационные порталы, что позволяло отказаться от этапа фильтрации документов (считалось, что все документы на этих сайтах относятся к заданной предметной области). Лингвистический анализ документов включал в себя сегментацию, извлечение наименований географических объектов, организаций и персоналий посредством библиотеки Polyglot [15], лемматизацию с помощью лингвистического анализатора Exactus [25] и извлечение дополнительных сущностей (занимаемые должности, страны, ресурсы, военные объекты, мероприятия) с помощью специализированного алгоритма. Этот алгоритм использует словари, лемматизацию и поиск подстрок с помощью алгоритма Ахо-Корасика [34]. Словари составлялись с частичным привлечением экспертов для задания эталонной лексики или поиска существующих ресурсов, подходящих для построения словарей. Некоторые словари (мероприятия, военные объекты, ресурсы, должности) были расширены схожими по контексту употребления словами с помощью алгоритма [35], обученного на Википедии. После тегирования лексических элементов выполнялась их фильтрация: лексемы, входящие в состав именованных сущностей, не входили отдельно в результирующее множество. После обработки каждый документ представлялся в виде множества упоминаний сущностей. Например, предложение «На строительство аэродрома, а также стационарных объектов частей и подразделений министром Минобороны РФ на остров Земля Александры планируется направить почти 7,19 млрд.руб.» преобразовывалось в «NOUN-рубль COUNTRY-Россия VERB-направить I-ORG-минобороны_рф POSITION-министр VERB-планироваться RES-остров NOUN-миллиард NOUN-строительство MIL_OBJ-подразделение MIL_OBJ-аэродром I-LOC-земля_александры NOUN-объект ADJ-стационарный».

В результате анализа имеющихся документов были получены выборки, состоящие из 13576 наборов (каждый из которых представляет элементы одного документа) суммарным объёмом более двух миллионов элементов (для русского языка) и 4315 наборов объёмом более 600 тысяч элементов (для английского языка).

По полученным наборам данных были построены тематические модели с детализацией в 2, 5, 10, 20, 50 и 100 тематик. Ниже приведены множества наиболее значимых ключевых слов тематик для русского языка и наиболее вероятные их интерпретации (детализация – 10 тематик):

– Добыча углеводородов в Ямало-ненецком автономном округе и на шельфе Северного ледовитого океана: *месторождение, компания, проект, нефть, газ, строительство, добыча, спг, миллион, газпром, договор, миллиард, тонна, янао, роснефть, шельф*.

– Транспортные вопросы, Северный морской путь: *морской, северный порт, путь, ледокол, судно, груз, навигация, река, миллион, рыба, атомный, Сабетта, Севморпуть, инфраструктура, флот, танкер*.

– Общественные мероприятия о поддержке северных регионов и коренных народов: *арктический, развитие, народ, международный, регион, коренной, форум, сотрудничество, совет, президент, губернатор, проблема, Россия, конференция, правительство, малочисленный*.

– Научные экспедиции, посвященные оценке масштабов изменения климата: *лед, ученый, станция, арктический, исследование, климат, полярный, изменение, температура, потепление, площадь, океан, таяние, экспедиция, глобальный*.

– Программы государственной поддержки и развитие инфраструктуры: *рубль, Якутия, чукотский, программа, правительство, округ, развитие, военно-транспортный самолёт, миллион, аэропорт, Чукотка, средства, работа, связь, бюджет, вертолёт, авиация, инфраструктура*.

– Развитие туризма на европейском севере: *проект, туризм, фестиваль, культура, конкурс, пройти, музей, мурманская область, международный, архангельская область, парк, выставка программа*.

– Морские научные экспедиции: *экспедиция, судно, северный, море, исследование, ледокол, работа, полюс, ученый, российский экипаж, земля, научно-исследовательский, Новая земля, Карский*.

– Проблемы защиты редких видов животных: *медведь, белый, животное, олень, вид,*

wwf, природа, человек, территория, популяция, морж, оленевод, птица.

– Развёртывание арктической группировки войск: *остров, северный флот, земля, сила, военный корабль, оборона, архипелаг, новый, пресс-служба, территория, учения, вооружение*.

– Политическое взаимодействие вокруг принадлежности шельфа: *шельф, российский, страна, северный морской, компания, США, освоение, проект, ресурс, море, Норвегия, международный, вопрос, развитие, нефть, совет, континентальный, граница, министр*.

Аналогичный список тематик для английского языка (детализация – 2 тематики):

– Добыча углеводородов, военное присутствие и северный морской путь: *oil russia company, drill, Norway,Barents, project, region, minister, take, part*.

– Вопросы экологии: *ice, polar climate, change, scientist, world, research, global, Jeffrey Gleason, bear, warm, climate change*.

Тематический анализ с другими уровнями детализации подтверждает адекватность разбиения информационного пространства на именно такие сюжеты (для русского языка). Объём собранной информации на английском языке оказался гораздо меньше, чем на русском, поэтому более детальный тематический анализ генерировал повторяющиеся сюжеты (добыча нефти, северный морской путь, военное присутствие, конфликт вокруг Arctic Sunrise). Таким образом, иностранные СМИ уделяли больше внимания конфликтным вопросам, а не развитию инфраструктуры, туризма и экологии, хотя, в целом, области внимания достаточно близки.

Помимо тематического анализа на основе полученных наборов была выполнена генерация ассоциативных правил с целью установления типов связей между сущностями, действий и т.п. Для этого применялся априорный алгоритм [26]. Настройка параметров (минимального покрытия и уверенности) выполнялась с помощью перебора значений по сетке с шагом, изменяющимся в геометрической прогрессии, начиная с 0.9 для покрытия и 0.9 для уверенности с шагом 0.5 и 0.7 соответственно. Перебор параметров заканчивался, либо когда достигались минимально допустимые значения покрытия (0.0001, что соответствует 10 документам для русского языка) или уверенности 0.55; либо когда количество сгенерированных правил превышало 1000. Всего по русскоязычным материалам было сгенерировано 2533 правила, а по англоязычным – 2832. Ниже приведены некоторые примеры сгенерированных правил:

– RES-нефть RES-шельф COUNTRY-Россия RES-углеводород RES-месторождение ~ EVENT-добыча.

– POSITION-ученый MIL_OBJ-судно EVENT-изучение ~ EVENT-экспедиция.

– RES-нефть RES-газ I-ORG-Газпром ~ EVENT-добыча RES-месторождение.

– MIL_OBJ-станция I-LOC-Мурманск ~ EVENT-экспедиция.

– EVENT-экспедиция RES-остров_врангеля POSITION-ученый COUNTRY-Россия ~ EVENT-исследование.

– MIL_OBJ-разведка I-ORG-Роснефть ~ RES-шельф.

– EVENT-встреча COUNTRY-США ~ COUNTRY-Россия.

– EVENT-exploration I-ORG-shell RES-climate I-LOC-Alaska ~ EVENT-drill.

– MIL_OBJ-ship EVENT-arrest EVENT-protest ~ I-ORG-Greenpeace.

– I-LOC-Murmansk EVENT-drill EVENT-protest ~ I-ORG-Greenpeace.

Такие правила можно использовать как для генерации шаблонов для дальнейшего детектирования ситуаций, так и для пополнения базы фактов. Это может осуществляться, например, путём добавления сущностей “Действие” для каждого глагола или элемента с тегом “EVENT” и связей к сущностям, соответствующим другим элементам правила. При этом правила имеет смысл генерировать за ограниченные периоды времени с перекрытием (скользящее окно).

Заключение

В статье выполнен аналитический обзор методов и систем информационной поддержки принятия решений на основе анализа информационного пространства по заданной проблеме. Рассмотрены методы сфокусированного сбора данных, методы извлечения информации из текстов на естественном языке, открытого извлечения информации и извлечения отношений. Предложен подход к разработке интегрированного метода и экспериментальных программных средств, опирающихся на методы

интеллектуального анализа данных и текстов для решения задачи информационной поддержки принятия решений, а также описана архитектура перспективных программных средств, реализующих этот подход.

Выполнен первичный анализ информационного пространства, цели которого успешно достигнуты. Жизнеспособность предлагаемой архитектуры и комбинации алгоритмов подтверждена. Выявлены основные актуальные сюжеты, некоторые ассоциативные правила, описывающие типичные ситуации.

В ходе дальнейших исследований планируется разработать экспериментальные программные средства информационной поддержки принятия решений, интегрирующие методы полнотекстового поиска и ситуативного анализа баз знаний. Также планируется разработать методику и провести полномасштабную экспериментальную проверку качества решения задачи ситуативного поиска с привлечением экспертов предметной области.

Литература

1. Imran M. et al. Processing social media messages in mass emergency: a survey //ACM Computing Surveys (CSUR). – 2015. – Т. 47. – №. 4. – С. 67.
2. Petrovic S. Real-time event detection in massive streams. – 2013.
3. Li R. et al. Tedas: A twitter-based event detection and analysis system // Data engineering (icde), 2012 ieee 28th international conference on. – IEEE, 2012. – С. 1273-1276.
4. Disaster SitRep – A vertical search engine and information analysis tool in disaster management domain / Li Zheng, Chao Shen, Liang Tang et al. // Proceedings of 2012 IEEE 13th International Conference on Information Reuse and Integration (IRI). — 2012. — P. 457–465.
5. Tweedr: Mining Twitter to inform disaster response / Zahra Ashktorab, Christopher Brown, Manojit Nandi, Aron Culotta // Proceedings of ISCRAM. — 2014. — P. 354–358.
6. Xiaohua L. et al. Recognizing named entities in tweets / Xiaohua Liu, Shaodian Zhang, Furu Wei, Ming Zhou // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies . – Association for Computational Linguistics. — 2011. — P. 359–367.
7. Bhattacharya A., Tiwari M. K., Harding J. A. A framework for ontology based decision support system for e-learning modules, business modeling and manufacturing systems // Journal of Intelligent Manufacturing. – 2012. – Т. 23. – №. 5. – С. 1763-1781.
8. Rao L., Mansingh G., Osei-Bryson K. M. Building ontology based knowledge maps to assist business process re-engineering //Decision Support Systems. – 2012. – Т. 52. – №. 3. – С. 577-589.
9. Hersovici M. et al. The shark-search algorithm. An application: tailored Web site mapping //Computer Networks and ISDN Systems. – 1998. – Т. 30. – №. 1. – С. 317-326.
10. Chen Z. et al. An improved shark-search algorithm based on multi-information // Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on. – IEEE, 2007. – Т. 4. – С. 659-658
11. Su C. et al. An efficient adaptive focused crawler based on ontology learning // Hybrid Intelligent Systems, 2005. HIS'05. Fifth International Conference on. – IEEE, 2005. – С. 6 pp.
12. Liu H., Janssen J., Milios E. Using HMM to learn user browsing patterns for focused web crawling // Data & Knowledge Engineering. — 2006. — Vol. 59, no. 2. — P. 270–291
13. Blanvillain O., Kasioumis N., Banos V. BlogForever Crawler: Techniques and Algorithms to Harvest Modern Weblogs //Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14). – ACM, 2014. – С. 7.
14. Florian R. et al. Named entity recognition through classifier combination //Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. – Association for Computational Linguistics, 2003. – С. 168-171.
15. Al-Rfou R. et al. Polyglot-NER: Massive multilingual named entity recognition //Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada. – 2015.
16. Wikipedia – свободная энциклопедия [Электронный ресурс] / Wikimedia. – URL: http://wikipedia.org (проверено 20.01.2016).
17. Bollacker K. et al. Freebase: a collaboratively created graph database for structuring human knowledge //Proceedings of the 2008 ACM SIGMOD international conference on Management of data. – ACM, 2008. – С. 1247-1250.
18. Manning C. D. et al. Introduction to information retrieval. – Cambridge : Cambridge university press, 2008. – Т. 1. – С. 496.
19. Соченков И. В., Суворов Р. Е. Сервисы полнотекстового поиска в информационно-аналитической системе (Часть 1) //Информационные технологии и вычислительные системы. М.: ИСА РАН. – 2013. – №. 2. – С. 69-78.
20. Takase S., Okazaki N., Inui K. Fast and Large-scale Unsupervised Relation Extraction. – 2015.
21. Angeli G., Premkumar M. J., Manning C. D. Leveraging Linguistic Structure For Open Domain Information Extraction //Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL. – 2015. – С. 26-31.
22. TAC Knowledge Base Population [Электронный ресурс] // NIST Information Technology Laboratory. – 2015. – URL: http://www.nist.gov/tac/2015/KBP/ (проверено 20.01.2016).
23. Hoffmann R. et al. Knowledge-based weak supervision for information extraction of overlapping relations //Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. – Association for Computational Linguistics, 2011. – С. 541-550.

24. Scrapy. A Fast and Powerful Scraping and Web Crawling Framework [Электронный ресурс] // Scrapy. – 2016. – URL: http://scrapy.org/ (проверено 20.01.2016).
25. Osipov G. et al. Relational-situational method for intelligent search and analysis of scientific publications //Proceedings of the Integrating IR Technologies for Professional Search Workshop. – 2013. – С. 57-64.
26. Agrawal R., Imielinski T., Swami A. Mining association rules between sets of items in large databases // Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data / ACM. — Vol. 22. — 1993. — P. 207–216.
27. Blei D. M. Probabilistic topic models // Communications of the ACM. – 2012. – Т. 55. – №. 4. – С. 77-84.
28. Д.А. Девяткин, Р.Е. Суворов, И.В. Соченков. Метод тематической кластеризации масштабных коллекций научно-технических документов // Информационные технологии и вычислительные системы. - 2013. - № 1. - С. 33-42.
29. Haklay M., Weber P. Openstreetmap: User-generated street maps //Pervasive Computing, IEEE. – 2008. – Т. 7. – №. 4. – С. 12-18.
30. Titan:Distributed Graph Database [Электронный ресурс] // DataStax. – 2016. – URL: http://thinkaurelius.github.io/titan/ (проверено 20.01.2016).
31. Lakshman A., Malik P. Cassandra: a decentralized structured storage system // ACM SIGOPS Operating Systems Review. – 2010. – Т. 44. – №. 2. – С. 35-40.
32. Joishi J., Sureka A. Vishleshan: performance comparison and programming process mining algorithms in graph-oriented and relational database query languages. – 2015.
33. Rodriguez M. A. The Gremlin graph traversal machine and language (invited talk) // Proceedings of the 15th Symposium on Database Programming Languages. – ACM, 2015. – С. 1-10.
34. Aho A. V., Corasick M. J. Efficient string matching: an aid to bibliographic search // Communications of the ACM. – 1975. – Т. 18. – №. 6. – С. 333-340.
35. Al-Rfou R., Perozzi B., Skiena S. Polyglot: Distributed word representations for multilingual nlp //arXiv preprint arXiv:1307.1662. – 2013.

Девяткин Дмитрий Алексеевич. Младший научный сотрудник ИСА ФИЦ ИУ РАН. Окончил Рыбинскую авиационную технологическую академию им. П.А. Соловьева в 2011 году. Автор 15 печатных работ. Область научных интересов: сфокусированный сбор данных, кластеризация текстов, машинное обучение. E-mail: devyatkin@isa.ru

Суворов Роман Евгеньевич. Младший научный сотрудник, аспирант ИСА ФИЦ ИУ РАН. Окончил Рыбинский государственный авиационно-технический университет им. П.А. Соловьёва 2012 году. Автор 15 печатных работ. Область научных интересов: анализ данных, машинное обучение, извлечение информации из текстов на естественном языке, высоконагруженные распределённые системы, графовые базы данных . E-mail: rsuvorov@isa.ru

Соченков Илья Владимирович. Научный сотрудник, заместитель заведующего лабораторией ИСА ФИЦ ИУ РАН, доцент кафедры информационных технологий РУДН. Окончил Российский университет дружбы народов в 2009 году. Кандидат физико-математических наук. Автор 50 печатных работ. Область научных интересов: интеллектуальные методы поиска и анализа информации, обработка больших массивов данных, защита сетей, контентная фильтрация, компьютерная лингвистика, распознавание образов. E-mail: ivsochenkov@gmail.com