Министерство образования и науки РФ
Московский педагогический гуманитарный университет
Имени В.И. Ленина
Факультет/институт Филологический

Магистерская диссертация

Тема: «Исследование лингво-статистических методов автоматического формирования ассоциативно-иерархического портрета предметной области»

Кафедра Компьюторная лингвистика

Выполнил:
Попова Ольга

Научный руководитель:
К.т.н., доцент
Шарнин М.М.

Рецензент:
К.фил.н., руководитель лаборатории
Компьютерной лингвистики и когнитивных
технологий обработки текстов
Козеренко Е.Б.

Москва
2016

Оглавление

Введение

Семантический анализ больших коллекций — ключевая проблема компьютерной лингвистики.. Тематическое моделирование — одно из современных приложений машинного обучения к анализу текстов, которое дает возможность определить тематическую структуру каждого документа из текстовой коллекции, а также тематический профиль всех слов этой коллекции. Современный анализ текстов практически всегда имеет дело с большими объемами данных, которые можно обработать лишь с помощью параллельных или распределенных реализаций алгоритмов тематического моделирования.

В рамках решения фундаментальной научной проблемы семантического моделирования опробована методика автоматизированного выявления ассоциативных и иерархических связей из интернет-текстов и выполнено построение ассоциативно-иерархических портрета по автономным необитаемым подводным аппаратам (АНПА).

Методика представляет собой единый итеративный алгоритм тематического поиска естестенно-языковых текстов по формирование большого корпуса текстов, выделение значимых словосочетаний (ЗС) и создание словарей ЗС, выявление иерархических и ассоциативных связей между ЗС создание словарей ЗС по ПО АНПА.

Работа базируется на гипотезе о привлечении ассоциативных связей для определения значений, полный смысл которых выявляется с помощью контекстных окружений, что дает возможность автоматизировать процесс разграничения значений. Значения слов и словосочетаний определяются векторными пространствами контекстных признаков, извлекаемых из текстов.

Сочетание методов Hierarchical Latent Dirichlet Allocation (hLDA), модели дистрибутивной семантики и модели семантических векторных пространств (СВП) предполагает улучшение параметров иерархической кластеризации терминов предметной области, связанных ассоциативными связями.

Методика и её программные средства позволяют автоматически выделять ключевые слова из ЕЯ-текстов, выявлять ассоциативные связи между ними, задавать иерархию терминов и ЗС методом кластеризации и строить словари ЗС различных предметных областей, а также отбирать материал для создания аналитического отчета по заданной области. В работе использовался улучшенный вариант метода LDA: впервые были построены темы, состоящие не только из набора слов, а также содержащие значимые словосочетания (ЗС), позволяющие учитывать порядок слов.

В отличие от хорошо известных методик, используемая в работе модель из изначального семантического контекстного пространства, не связанного с конкретной тематикой, автоматически выбирает ту или иную предметную область и её компоненты: значимые словосочетаний (ЗС), ассоциативные связи и, соответственно, контексты для их выделения. В результате строится система множественных ассоциативных связей, а затем формируется ассоциативно-иерархический портрет предметной области – АИППО. Ассоциативно-иерархический портрет и онтология взаимно дополняют и обогащают друг друга. Терминология и иерархические связи, заданные в онтологии, служат входными данными для автоматического построения АИППО, который в свою очередь дополняет онтологию найденной в Интернете актуальной расширенной лексикой и связями. В результате получается новая мощная комбинация, превышающая возможности стандартной онтологии по составу и актуальности лексики. Множество иерархических связей АИППО вместе образуют классификатор терминов, который помогает в поиске и навигации по терминам предметной области. Классификатор делит предметную область на части, что создает систематичность ее исследования и разработки.

Для достижения цели проекта необходимо решить задачи разработки комплексной технологии по обработке больших массивов интернет-текстов с целью извлечения и систематизации знаний. Технология заключается в анализе текстов естественного языка, выявлением ассоциативных и иерархических

связей и построении на этой основе АИППО (ассоциативно-иерархического портрета ПО).

Для сбора больших объемов текстовых данных из Интернет по заданной предметной области (в частности, АНПА) были использованы технологии семантического серфинга.

Решена задача разработки методов, позволяющих автоматически создавать и использовать АИППО. В связи с этим трудоемкость разработки АИППО стала на несколько порядков ниже разработки традиционных онтологий и тезаурусов. Это позволяет АИППО стать реальным дополнением онтологий и тезаурусов в новом поколении интеллектуальных Интернет-технологий, которые отличаются гораздо большей полнотой охвата терминологии. Кроме того, возможна демонстрация возможностей АИППО для мониторинга предметной области АНПА.

В рамках работы также проведен экспермент по использованию рассматриваемой методики для построения иерархии понятий по предметной области АНПА. Полученная иерархия может служить основой онтологии данной предметной области. А результате эксперимента методика подтвердила свою эффективность при автоматизированном построении онтологий различных предметных областей.

1. Методы, используемые для создания базы релевантных интернет-текстов

Корпусы (сборники образцов языка, поизведенные в натуральнох контекстах и без экспериментального вмешательства) играют все более центральную роль в различных отраслях лингвистики и смежных дисциплин. Например, корпусы ключительно используется для приведения фактического доказательства в

вопросах теоретическойи прикладной лингвистики ([21]), при моделировании процессов овладения языком ([10]), в лексикографии ([24]) , а также в большом числе задач по обработке естественного языка ([20]).

Исследователи все больше заинтересованы в Сети как потенциальном источнике лингвистических данных ([18]). Веб содержит огромное количество текстовых данных для постоянно возрастающего числа языков, он содержит много различных жанров и специализированных текстов, и, конечно же , он является и будет являться"возобновляемым" источником языка, до тех пор , пока люди отправляют туда новые данные.

Стоит сначала рассмотреть некоторые общие плюсы и минусы использования Интернета в качестве корпуса, не утверждая , что ни один из минусов не является специфическим для Web корпусов сам по себе (раздел 1.1). Мы кратко просмотрим основные подходы , которые были приняты на вооружение для получения лингвистических данных из Интернета, настаивая на лингвистически нацеленном кроулинге, как единственном жизнеспособном долгосрочном решении (раздел 1.3).

1.1.Плюсы и минусы использования Интернета в качестве корпуса

Одним из основных преимуществ создания веб-КОРПУСА это размер. Размер дает возможность использовать более простые алгоритмы обработки данных, которые тем не менее показывают результаты не хуже при обучении на большем количестве данных, чем более сложные алгоритмы (так, в авторитетной работе([3]), Банко и Brill показали , что даже простой алгоритм устранения неоднозначности(disambiguation algorithm) тем не менее превосходит более сложные методы , когда он обучается на большем количестве данных)). Кроме того, WEB дает возможность проводить более полные и точные исследования любых явлений. Например, Майр ([19]) показал , что веб, в отличие от БНК, достаточно большой , чтобы позволить полное

исследование грамматикализации.. Терни ([27]) показал , как простой алгоритм , опираясь на веб - частоты обгоняет намного более сложный метод , обучающийся на корпусах меньшего размера в задаче обнаружения синонимов . еще одно преимущество web- это то , что он позволяет быстро и дешево строить корпуса на многих языках . Результаты [23] и [28] показывают , что Web корпуса, построенные одним исследователем буквально в течение нескольких минут, являются, с точки зрения разнообразия жанров, тем и лексикона, ближе к традиционным "сбалансированным" корпусам , таким как BNC , чем к моноисточниковым корпусам, таким как корпуса на основе лент новостей. Кроме того, эти корпуса имеют тенденцию отражать более поздние языковые периоды, чем традиционные корпуса, которые часто подвергаются определенной задержке между временем производства материалов , которые в конечном итоге становятся корпусом и временем публикации самого корпуса. Третим преимуществом веб - корпусов является то , что они потенциально могут содержать количество жанров , которые не присутствуют в традиционных письменных источниках. Такие явления, как ведение блога blogging должно представлять интерес для лингвистов , так как они создают огромное количество письменных образцов на огромное разнообразие тем, которые спонтанно производствводятся непрофессиональными писателями. Более того, Сеть предоставляет множество образцов интерактивного общения , которое, хотя и в письменной форме, обладает некоторыми характеристиками устного общения ([25]). В то же время, так как Использование Интернета для различных архивных и практических целей расширяется, трудно думать о традиционных письменных жанрах, не представленных в Интернете. Кроме того, Вев дает возможность быстро обнаружить последнии новации , относящиеся у различным областям науки и производства.

Конечно, Web корпусы, создают также некоторые проблемы. Первое, такие корпусы , как правило, содержат много шума, такой как автоматически сгенерирированный неязыковой материал и дублированные документы. Хотя,

сбор корпуса в 1 млрд слов из нескольких источников, не имеющих отношения к Web, в течение нескольких дней, вероятно , не представляется возможным, но, если бы это было возможно, полученный корпус почти наверняка имел бы точно те же проблемы шума. И нет никаких оснований полагать ,что корпус будет менее чистый, а его содержание менее контролируемым Кроме того, если исследователь планирует распространить большой Веб – корпус, изготовленный из миллиона документов, достаточно сложным может стать процесс получения разрешения на использование документов от всех владельцев авторских прав. Хотя этих проблем не избежать и при других способах построения корпусов. В любом случае, не правильно было бы ссылаться на вышеперечисленные проблемы как на проблемы Web корпуса, скорее, они являются проблемами больших корпусов, построенных в короткие сроки и с небольшими ресурсами, и несомненно они появляются с помощью веб - корпусов , так как веб - позволяет построить "быстрый и грязный" большой корпус.

В то время как, в общем, мы не видим какие - либо минусы , которые являются уникальными для веб - корпусов, есть, тем не менее, проблемы , характерные в частности для Web-а-корпусной методологии, а именно те, которые в значительной степени полагаются на Google или другие коммерческ поисковые системы для получения веб - данных.

1.2.Три подхода к Сети как к лингвистическому корпусу

1.2.1. Оценка частот с помощью поисковых систем

Вероятно, самый старый и самый распространенный подход при использовании Веб для лингвистических целей - это выдача запроса на поиск к поисковой машине, такой, как Google, и использование "хит-счетчика", сообщаемого с помощью поисковой системы в качестве оценки частоты появления искомой

строки на целевом языке. Эта стратегия оказалась весьма успешной при решении различных задач. Ссылаясь только на один известный пример, Терни ([27]) показал , что простой подход , основанный на сборе запросов хит подсчетов типа A NEAR B в поисковой системе AltaVista при решении задачи обнаружения синонимов гораздо эффективнее, чем более изощренными методы, например, латентный семантический анализ, при использовании традиционного корпуса в качестве входных данных.

Однако, несмотря на эмпирические успехи, этот подход очень проблематичен. Во- первых, типы запросов для поисковой системы очень ограничены: например, нельзя ограничить поиск на основе частей речи, нельзя использовать регулярные выражения. Действительно, так как пользователи поисковой системы , как правило , интересуются тем, к чему относятся их условия поиска, а не их языковыми свойствами, поисковые системы как правило, выполняют ряд упорядовачинаний для поисковых запросов, которые могут быть чрезвычайно раздражающими для лингвистов, таких как игнорирование падежа, отсечение диакритических знаков, игнорирование апострофа и дефиса. По тем же причинам, поисковые системы часто игнорируют или производят очень странные результаты при запросе служебных слов. Таким образом, типы вопросов исследования , которые можно достичь с помощью этой методологии априори весьма ограничены: можно использовать счетчики поисковой системы для поиска частоты предварительно отборанного контента слова и словосочетания, только если точное соответствие не имеет решающего значения для исследования вопроса.

Во- вторых, поисковые компании, по понятным причинам, не публикуют подробную информацию о том , как они собирают, и индекссируют и возвращают результаты запроса, а также услуги , которые они обеспечивают, часто непредсказуемо обновляющиеся после технологического и изменения на рынке, как правило, изменчивы и непостоянны. Так, в апреле 2004 года,

AltaVista вдруг перестала поддерживать NEAR-оператора, сделав оригинальный алгоритм Терни невозможным для использования.

В общем , использование хитов в поисковых системах подходит кажется только для экспериментальных исследований, или очень ограниченных контекстов, но не годятся для долгосрочного подхода к использованию Веб в качестве корпуса.

1.2.2. Построение корпусов с помощью запросов поисковых машин

Вместо того чтобы полагаться на подсчеты поисковых систем, можно оформить автоматизированные запросы к поисковой системе (поисковые системы , такие, как Google и Yahoo! предоставляют API для Веб-сервисов, которые позволяют пользователям выполнять определенное количество автоматизированных запросов в день), восстановить страницы , возвращаемые с помощью поисковой системы, и обрабатывать их , чтобы построить корпус.

Такой подход был изучен различными учеными, в том числе [15, 5, 14], и очень обширно Sharoff ([23]) и Ueyama ([28]). Sharoff показывает, как корпусы на разных языках, построенные путем выдачи запросов для случайных комбинаций частых слов в поисковой системе Google и извлечение и обработка страниц, найденных таким образом, имеют характертики, больше похожие на характеристики сбалансированного корпуса типа BNC, чем на характеристики моно-источниковых корпусов. С другой стороны, запросы с различными вариантами терминов ведут к корпусам с очень разным составом. В частности, с использование основных словарных слов в качестве условий запроса приводит к корпусам, которые характеризуются высокой долей текстов с бытовой тематикой, производенными непрофессиональными писателями, в то время как термины для запросов, отбираемых из более формальных источников, приводят к созданию корпусов, которые характеризуются

государственными научными / техническими темами и профессиональными авторами.

Такой подход не так сильно зависит от Google (или другой поисковые системы), чем тот , который обсуждался выше (п. 3.1). Google используется для получения списка документов, но потом эти документы извлекаются и подвергаются пост-обработке исследователем (токенизация , POS-тегирование, и т.д..). после этого стабильность данных больше не будет зависеть от Google, а исследователь имеет полный доступ к корпусу.

Однако, здесь есть свои проблемы. Набор страниц, которые извлекаются по - прежнему зависит от поисковой машины. Кроме того, по очевидным причинам поисковые машины ограничивают количество данных, которые могут быть получены с помощью автоматизированно выполненных запросов (например, в случае Google можно максимально получить 10K URL - адресов в день - и, конечно, не все извлекаемые страницы подходят для корпусов. Таким образом, в то время как огромное количество доступных здесь данных - один из главных факторов, привлекающих лингвистов использовать Веб, построение действительно большого корпуса (порядка миллионов документов) с помощью этого метода весьма спорно.

1.3.Лингвистический кроулинг Веба.

Похоже, что единственный жизнеспособный долгосрочный подход к построению Веб-корпусов для лингвистов - выполнить свои собственный кроулинг Интернета. Это дает лингвистам возможность быть полностью независимыми от коммерческих поисковых систем, и обеспечивает полный контроль в течение всей процедуры строительства корпуса. Однако это и самый трудный подход для реализации, особенно если цель - большой корпус. Необходимы значительные вычислительные ресурсы, чтобы провести широкомасштабный кроулинг. Кроме того требуется «очистка» произведенных

кроулингом данных. В зависимости от дальнейшего использования коллекции к ней предъявляются определенные требования, накладывающие ограничения на ее качественные характеристики. Такими требованиями, например, могут быть определенный уровень релевантности тематике, отсутствие рекламы и текстов-дубликатов и др. Кроме того может потребоваться удаление страниц нецелевого языка или другого рода страниц, которые желательно исключить по тем или иныи причинам; отсечение HTML-кодов и "шаблоннов". Эти шаги, включая релевантность и удаление дублирования используется в технологии Keywen. Затем, с учетом целей работ, может потребоваться хотя бы минимальное аннотирование данных POS - тегами и Леммами (адаптируя аннотирования к языку Веб$_a$). - . Когда данные, полученные с помощью кроулинга, порядка сотен гигабайт, все эти шаги становятся далеко не тривиальными,и не только с точки зрения лингвистического качества результатов, но также с точки зрения времени и эффективности.

В самом деле, насколько похоже, сто ни один из существующих проектов больших лингвистических кроулингов Сети не прошел через все этапы описанной процедуры. Терабайтный Corpus , описанный в работе [11], не подвергался ни фильтрации языка , ни какой другой форме пост-обработки, и не анотирован никакой лингвистической информацией. Корпус английского Академического веб-сайта [26] также не подвергался никакой какой постобработке, кроме простой токенизации. Китайский корпус [13] имеет фильтрацию с точки зрения обнаружения языка, но это все.

Еще одна большая работа в этой области проводится группой лингвистов в рамках WaCky проекта[ ]. Были построены крупные (> 1 миллиард токенов) Веб-извлеченныt корпуса немецкого и Итальянского языков, которые затем тщательно были обработаны и аннотированы основной морфологической информацией.

1.4. Построение большого Веб- корпуса предметной области (АНПА)

Целью первоначального этапа работы является создание тематической текстовой коллекции, т. е. массива текстов, относящихся к определенной тематике. Такие коллекции могут играть как чисто информационную роль, так и использоваться при решении широкого класса задач (машинное обучение(алгоритмов), построениеи наполнение онтологий, составление различного вида словарей-морфологических, синонимов, тезаурусов и т. п.- и рефератов, кластеризация(например, для выявления трендов в за-данных областях знаний) и т. д.

Рассматриваемая здесь методика построения корпусов по конкретной предметной области «АНПА» реализуется на сверхбольших объемах интернет-данных, так называемых ‘big data’, свободно представленных в среде Интернета. Обработка больших массивов текстов из Интернета позволяет собирать необходимые статистические данные для формирования достаточно полной картины о ПО, представленной в виде АИППО. Использовались методы сбора больших объемов текстовых данных из интернет по заданной предметной области (АНПА) при помощи технологии семантического серфинга, впервые примененной при построении системы Keywen. Технология семантического серфинга способствует повышению полноты и точности результатов, так как она использует точное задание предметной области в виде большого набора ключевых терминов, а также проверку релевантности найденных документов. Были реализованы методы итеративного формирования корпусов текстов в заданной предметной области, на базе которых создаются корпуса текстов на естественном языке (ЕЯ-тексты), постоянно пополняемые из Интернет. Было выделено более 400 ключевых терминов (ключевых слов и значимых словосочетаний).

Работа осуществлялась методами автоматического интернет-серфинга в результате круглосуточного функционирования двух серверов и методом

круассординга силами авторов рассматриваемого проекта. В качестве источников данных при решении этой задачи используются интернет-ресурсы. Для их сбора реализуется механизм метапоиска и применяются универсальные методы, ориентированные на произвольный формат данных. для окончательного формирования тематической коллекции предметной области. Предлагаемая система предназначена для автоматического формирования тематических текстовых коллекций на основе интернет-ресурсов, с дальнейшем решением задач по созданию ассоциативно-иерархического портрета предметной области (АИППО).

Система состоит из нескольких взаимосвязанных компонентов, которые объединяются в два основных блока: блок сбора информации и блок обработки текстов. Для сбора текстов используются механизмы метапоиска. На вход подается поисковый запрос, содержащий набор ключевых терминов, характеризующих тематику коллекции. Поисковый запрос формируется на основе множества ключевых слов и значимых словосочетаний. Осуществляется пребор всех комбинаций ключевых терминов (единичные термины, пары трерминов, тройки терминов и т.д.), так, чтобы длина запроса не превышала определенной границы N. Величина N определялась эксперементально, исходя из анализа эффективности поисковых запросов – слишком длинные запросы выдают пустые результаты. На начальном этапе запрос задается непосредственно пользователем с учетом предварительных знаний о исследуемой ПО и впоследствии генерируется самой системой в виде URL ссылок. Далее запрос передается компонентам, отвечающим за сбор информации, работающим с интернет-ресурсами общего вида и использующим механизм метапоиска для сбора данных. Затем запрос передается подключенным к системе информационно-поисковым системам (ИПС) – GOOGLE, YANDEX и т.д.

1.4.1. Преимущества собственного поискового агента перед готовыми системами. Преимущества использования нескольких ИПС.

Для поиска по всей сети Интернет создан собственный поисковый агент (Kewen), обходящий страницы Интернета и составляющий поисковый индекс. Конечно, можно использовать результаты поиска существующих информационно-поисковых систем. Создание собственного поискового агента привлекательно тем, что алгоритм его работы открыт и известен, в то время как поисковый алгоритм большинства ИПС является коммерческой тайной и старательно оберегается их создателями. Однако сканирование сети Интернет – достаточно трудоемкий процесс и требует больших временных и аппаратных затрат, к тому же полнота поиска все равно будет хуже, чем в метапоисковой системе, использующей несколько ИПС, покрывающих различные участки Интернета. Помимо этого каждая ИПС имеет свой алгоритм ранжирования результатов поиска, из-за чего поисковая выдача разных ИПС может сильно различаться. Поэтому метапоисковая система, использующая несколько ИПС, может вернуть более точный и полный список результатов на запрос пользователя. Основная идея метапоиска заключается в формировании поисковой выдачи (результатов поиска) за счет объединения и переупорядочивания поисковых выдач всех задействованных ИПС. Метапоисковые системы в отличие от традиционных ИПС обычно не имеют собственного поискового индекса. Они обрабатывают поисковый запрос пользователя и отправляют его нескольким ИПС, которые либо выбираются из списка ИПС автоматически, либо задаются пользователем вручную. Ответ каждой ИПС обрабатывается, и результаты поиска объединяются в единый список[2].

1.4.2. Работа метопоисковой машины

В работе метапоисковой машины можно выделить три этапа:

- предобработку запроса (Предобработка запроса заключается в его расширении с использованием словарей или тезаурусов);
- выполнение запроса несколькими ИПС,
- постобработку результатов (В постобработку входит приведение результатов к единому виду, удаление повторяющихся ссылок, фильтрация);
- загрузка текстов для создание текстовой коллекции.

Также в метапоисковых системах большое значение имеет объединение и переупорядочивание результатов, полученных от разных ИПС. Однако для задачи сбора текстов порядок следования результатов не так важен, как для метапоисковых систем, предоставляющих результат пользователю, который хочет получить наиболее релевантные результаты в начале списка.

1.4.3. Расширение поискового запроса

Взаимодействие с поисковыми системами. Для взаимодействия с ИПС используются их поисковые программные интерфейсы (API). Такие ИПС, как Google, Яндекс, Yahoo!, Bing позволяют отправлять запросы к их поисковым базам и получать ответы либо в формателXML, либо JSON (в бесплатных версияхAPI существуют ограничения на число запросов). Например, чтобы воспользоваться Яндекс.XML-1 (поисковыйAPI, предоставляемый Яндексом), нужно отправить на специальный адрес запрос (либо POST, либоGET), в котором в качестве параметров указывается поисковый запрос, правила сортировки и фильтрации, число требуемых результатов. В ответ возвращается список ссылок на страницы, которые Яндекс считает релевантными исходному запросу. При этом для каждой страницы выдается заголовок, кодировка, дата и время изменения, тип документа и сниппет– небольшой отрывок текста из найденной страницы, как правило содержащий контекст, в котором встретилось ключевое слово.

1.4.4. Возвращение результатов запроса ИПС

После того, как ИПС вернули результаты обработки запроса, последние (результаты) объединяются, из них удаляются одинаковые ссылки и формируется единый список ссылок на HTML-страницы с некоторой метаинформацией: время получения ссылки, время создания страницы (если известно), заголовок и сниппет. Обычно все ИПС предоставляют эти данные. Затем происходит загрузка страниц по ссылкам из списка. Основная задача этого этапа состоит в том, чтобы загрузить HTML-страницу по ссылке, используя HTTP-протокол, определить кодировку страницы и преобразовать ее в формат UTF-8

Специфика получаемых данных. В число возвращаемых ИПС результатов могут входить не только ссылки на HTML-страницы, но и на документы различных форматов (например, .doc, .pdf). На данном этапе разработки системы рассматриваются только HTML-страницы, а документы остальных типов игнорируются.

1.4.5. Вид данных в интернете

В настоящее время большая часть информации в сети Интернет представлена в виде HTML-страниц. Основной его (формата) недостаток, затрудняющий автоматическое извлечение информации, состоит в том, что в HTML не разделяются данные и их представление. Теги в HTML не имеют семантического значения, один и тот же тег может использоваться как для выделения блока информации, так и для создания навигации по сайту. В стандарте HTML-5 были добавлены новые теги, имеющие семантическое значение[7]. Одним из них является тег<article>, используемый для выделения независимой части документа или сайта (например, статья в блоге). Применение семантических тегов значительно бы облегчило автоматическую обработку сайтов, но далеко не все создатели сайтов применяют новые

семантические теги, даже если знают о них. Интернет – это хаотически развивающаяся система, поэтому приходится учитывать все разнообразие составляющих его сайтов при их автоматической обработке и извлечении из них информации.

1.4.6. Извлечение текстов в HTML странице .

В типичной HTML-странице можно выделить навигационную часть и содержательную часть. К навигационной части можно отнести меню, «шапку страницы», «низ страницы», элементы оформления. Обычно навигационная часть одинаковая у всех страниц одного сайта. Содержательная часть – основной контент страницы, то, ради чего она была создана.

Много исследований посвящено извлечению содержательной информации из HTML-страницы.

Можно выделить два направления разработок:

- анализ структуры отдельной HTML-страницы[8; 9];
- анализ нескольких страниц, выделение повторяющихся элементов[10; 11].

Подход, основанный на анализе множества страниц, базируется на предположении, что у всех страниц с одного сайта элементы оформления одинаковые, но содержательная часть разная. Значит, путем сравнения некоторого количества таких страниц можно выделить повторяющиеся части и удалить их, признав элементами оформления и навигации. То, что останется, и будет содержательной частью. В этом состоит основная идея данного подхода. Например, в работе[11] каждая страница разделяется на блоки, и после сравнения нескольких страниц те блоки, которые оказались уникальными, считаются содержательной частью. Выделение повторяющихся элементов требует наличия большого количества структурно похожих страниц (с одного сайта), в нашем же случае страницы

в поисковой выдаче достаточно разнообразные, что делает невозможным применение данного метода в рамках поставленной задачи. Если нет возможности выполнить разбор нескольких страниц, принадлежащих одному сайту, применяются методы, которые анализируют структуру отдельной страницы и на основе этого анализа делают выводы о ее содержимом.

1.5.Алгоритм построения Веб-корпуса по технологии Keywen

Система построения Веб-корпуса по технологии Keywen называется также системой семантического серфинга. Построение Веб-корпуса начинается с задания списка ключевых терминов и автоматических их запросов в различных поисковых машинах. В результате обработки запросов от поисковых систем получается множество M_1 текстовых документов. Это множество просматривается на предмет выделения URL-ссылок, по выбранным ссылкам формируется расширенное множество документов M_2. Множество M_2, в свою очередь, тоже содержит URL-ссылки, потому процесс расширения множества документов можно продолжать и далее, несколькими итерациями. При этом необходимо проверять вновь найденные документы на наличие в них первоначальных ключевых терминов. Эта проверка осуществляется вручную. В этом случае и последующие итерации вносят существенный вклад в пополнение множества искомых документов и списка ключевых слов. Далее, документы из множества M_2 делятся на предложения (или на фрагменты, близкие по длине к обычным предложениям). В результате составляется база данных с записями вида:

2 Методы, используемые для анализа текстов ЕЯ ПО

2.1 Выделение ключевых слов из текста

Извлечение ключевых терминов является базисным этапом для многих задач обработки естественного языка, таких как классификация и кластеризация документов, суммаризация текста и выявление общей темы документа, а также информационный поиск, электронный документооборот, лингвистика, мониторинг бизнес-процессов, проведение научных исследований, библиотечные и патентные технологии и др.. Объемы и динамика информации в этих областях, делают актуальной задачу автоматического выделения ключевых слов и фраз. Эти слова и словосочетания могут использоваться для создания и совершенствования терминологических ресурсов, а также для эффективной обработки документов в информационно-поисковых системах (индексирования, реферирования и классификации). Необходимость обобщения и систематизации исследуемых и разрабатываемых методов извлечения ключевых слов из текста, их классификации и практического применения существовала всегда.

В последние годы разработано множество подходов автоматизированного извлечения ключевых слов и словосочетаний, в основе которых лежат лингвистические, статистические и спектральные методы. Их анализ показывает, что создание эффективных экстракторов будет и далее являться одним из важнейших трендов компьютерной обработки текстов.

Преимуществами статических методов являются простота реализации и удовлетворительное качество работы, когда обучающее множество – коллекция документов для сборки статистики – удачно подобрана, недостатком - зависимость работы методов от обучающего множества.

Преимуществами семантических методов, во-первых, является то, что эти методы не требуют обучения, что важно, если нет достаточно

качественной обучающей коллекции. Во-вторых, они позволяют группировать термины в семантически близкие сообщества и устанавливать связи между ними, что необходимо при решении задач интеллектуального индексирования и поиска. К недостаткам относится сложность снятия омонимии и установления отношений между терминами, связанная со сложной структурой естественных языков, что, зачастую, является главной причиной лишь незначительного повышения эффективности поиска.

Таким образом, вероятно, наиболее разумным решением является комбинирование методов различной природы. Поиск наиболее удачного сочетания имеющихся в настоящий момент ресурсов и методов – важная задача для дальнейших исследований.

2.1.1 Общая схема извлечения ключевых слов из текста

Общая схема извлечения ключевых слов из текста практически одинакова для всех используемых методов и состоит из следующих шагов:

1. Предварительная обработка текста. Исключение элементов маркировки, приведение слова к словарной форме, удаление стоп-слов, не несущих смысловой нагрузки (предлоги, союзы, частицы, местоимения, междометия и т. д.) [4], [22].
2. Отбор кандидатов в ключевые слова.
3. Фильтрация кандидатов в ключевые слова (анализ значимых признаков для каждого кандидата).
4. Отбор ключевых слов из числа кандидатов.

Несколько примеров реализации некоторых шагов указанной схемы.

В [22] авторы обосновывают универсальность алгоритма извлечения ключевых слов и показывают различия методик на уровне процедур обработки текста и необходимых для них лингвистических знаний. Так, фильтрация кандидатов в ключевые слова производится в зависимости от метода извлечения ключевых слов. Например, при статистическом методе, фильтрация заключается в отборе определенного количества наиболее частотных лексем. Согласно схеме, приведенной в [3], кандидаты в ключевые слова фильтруются по статистическим и лингвистическим критериям. Далее для каждого кандидата формируется вектор в пространстве признаков. Процедура завершается сортировкой

всех кандидатов по вероятности быть ключевым словом и отбором заранее определенного числа кандидатов.

Отбор ключевых слов из числа кандидатов включает в себя расчет весов их информативности, который позволяет оценить их значимость по отношению друг к другу. Здесь стоит упомянуть известную метрику TF-IDF [14].

2.1.2 Классификация методов выделения ключевых слов

Извлечение ключевых терминов является базисным этапом для многих задач обработки естественного языка, таких как классификация и кластеризация документов, суммаризация текста и выявление общей темы документа.

Как было отмечено выше, в основе большого многообразия подходов автоматизированного извлечения ключевых слов и словосочетаний лежат лингвистические, статистические и спектральные методы.

В этой связи хотелось бы отметить два направления, которые очевидно имеют особую актуальность. Во-первых, анализ социально-экономических явлений и процессов, основанный на обработке тестов динамических ресурсов Интернет, главным образом социальных сетей. И во-вторых, выявление новых направлений в научной и инновационной деятельности, основанное на обработке текстов в специализированных хранилищах научно-технической литературы.

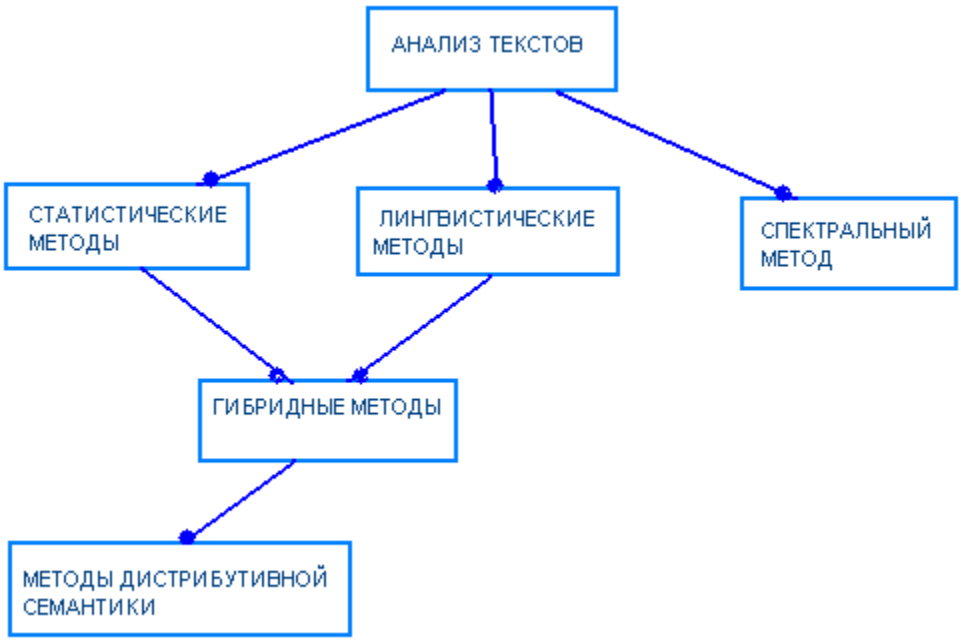
В работе [20] говорится о лингвистических и статистических критериях выделения ключевых слов. Статистические методы основаны на частоте употребления терминов, лингвистические критерии учитывают в первую очередь типичную структуру именных терминологических словосочетаний. Но лучшее качество, по мнению авторов, достигается при комбинации лингвистических и статистических подходов. Авторы статьи [22] считают, что основные типы методов и моделей автоматического извлечения ключевых слов можно разделить на статистические и гибридные (то есть включающие как статистические, так и лингвистические методы). Отдельно выделяются методы, требующие и не требующие наличия корпуса текстов одной тематики. Различие методик, по их мнению, определяется процедурами обработки текста на каждом из этапов и количеством необходимых для этих процедур лингвистических знаний.

В [11] предлагается спектральный метод поиска ключевых слов в полнотекстовых базах данных, основанный на теории спектрального анализа и использовании алгоритма быстрого преобразования Фурье.

В настоящее время широкое распространение получила разработка программных систем автоматического извлечения ключевых слов. Описание и сравнение наиболее популярных в технологии автоматического выделения ключевых слов, таких, как OpenCalais,

Extractor, Yahoo! Term Extraction Web Service, TerMine, Maui, TextAnalyst, AOT, ContentAnalyzer, Семантическое зеркало и других, можно найти в [2]. Рассматривая и сравнивая эффективность этих систем, авторы делают выводы о недостаточном качестве их работы. Отмечается, что не все системы поддерживают русский язык.

Таким образом, существуеь множество способов автоматизированного извлечения ключевых слови фраз из текста, и в основе всех этих способов лежат, как уже упоминалось, лингвистические, статистические, гибридные методы, а также спектральные методы, которые можно выделить в отдельную группу.



Далее каждая из этих групп методов можно рассмотреть более подробно.

2.1.2.1 Статистические методы

Статистические методы основываются на численных данных о встречаемости слова в тексте.

Классическими подходами в этой области считаются метод TF-IDF, используемый для выделения ключевых слов, и анализ коллокаций – для выделения словосочетаний.

TF-IDF (term frequency-invert document frequency) - статическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. В результате его применения больший вес получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употребления в других документах.

Для выделения многословных терминов используется анализ коллокаций, где коллокация – словосочетание, состоящее из двух или более слов, имеющее признаки синтаксически и семантически целостной единицы, в котором выбор одного из компонентов осуществляется по смыслу, а выбор второго зависит от выбора первого. В отличие от свободного словосочетания, коллокация определяет, какие слова могут быть использованы вместе.

Коллокации выявляются при лексическом анализе текста. Для этого используются различные меры ассоциативной связи, которые оценивают, является ли взаимное появление лексических единиц случайным, или оно статически значимо. [Новикова Автоматическое выделение терминов из текстов предметных областей и установление связей между ними]

В [22] отмечается, что преимуществами статистических методов являются универсальность алгоритмов извлечения ключевых слов, отсутствие необходимости в трудоемких процедурах построения

лингвистических баз знаний, простота реализации. К сожалению, удовлетворительного качества результатов статистические методы часто не обеспечивают. Кроме того, область эффективного применения статистических моделей хороша для языков с бедной морфологией; в случаях же естественных языков с богатой морфологией, в частности, для русского языка, могут возникнуть проблемы.

2.1.2.1.1 Методы выделения ключевых слов

Наиболее простой статистический метод из-влечения ключевых слов предполагает построение множества кандидатов ключевых слов путем ранжирования всех словоформ или лексем документа по частоте. Фильтрация заключается в отборе в качестве ключевых определенного количества наиболее частотных лексем. Этот метод является первым методом автоматического извлечения ключевых слов. Он разрабатывался, например, в работах Г.П. Луна [Luhn H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development. 1957, vol. 1, no. 4, pp. 309–317.], Р.Г. Пиотровского [Пиотровский Р.Г., Бектаев К.Б., Пиотров-ская А.А. Математическая лингвистика: учеб. по-собие для пед. институтов. М.: Высшая школа, 1977. 383 с. [Piotrovskiy R.G., Bektaev K.B., Piotrovskaya A.A. Matematicheskaya lingvistika. (Mathematical Linguistics). Moscow, Vysshaya shkola, 1977, 383 p.]4.] и широко используется до сих пор. Распространенность метода отбора ключевых слов исключительно на основе частот лексем объясняется его простотой.

При использовании частоты слова в документе в качестве единственного параметра для автоматического извлечения ключевых слов подсчет общей частоты словоформ из парадигмы одной лексемы чаще всего осуществляется следующим образом: общая частота ключевых слов подсчитывается путем сравнения словоформ, нормализованных к одной форме, как правило, к основе или лемме. Автоматическая нормализация словоформы по сути дела представляет собой задачу морфологического анализа и достаточно проблематична сама по себе.

При статистических подходах к извлечению ключевых слов используются простые эвристические алгоритмы, чаще всего нормализующие словоформу к ее квазиоснове, отсекая от словоформы определенное количество букв. Такие алгоритмы называют стемминг-алгоритмами, наиболее известным из которых является темминг-алгоритм Портера [Porter M.F. An Algorithm for Suffix Stripping. Readings in Information Retrieval. Morgan Kaufmann Publishers Inc., 1997, pp. 313–316.]. Нормализованные словоформы ранжируются по частоте и те из них, чья частота выше заданного порога, считаются ключевыми. вые слова, как правило, выдаются в усеченном виде квази-основ. Статистические методы извлечения многокомпонентных ключевых слов в качестве необходимого этапа построения множества кандидатов включают вычисление n-грам [Jiao H. Chinese Keyword Extraction Based on N-Gram and Word Co-occurrence. Proceeding CISW '07 Proceedings of the 2007 International Conference on Computational Intelligence and Security Workshops. Harbin, 2007. pp. 152–155., 34 ; Sarkar, K. An N-Gram Based Method for Bengali Keyphrase Extraction / K. Sarkar // Information Systems for Indian Languages. Springer Berlin Heidelberg, 2011, pp. 36–41]. С одной стороны, частота употребления слова несомненно арактеризует важность слова для данного документа, но, с другой стороны, ключевые слова, как подчеркивали исследователи руппы «Статистика речи» Р.Г. Пиотровского, и другие, не всегда являются самыми частотными [Алексеев П.М., Герман-Прозорова Л.П., Пиотровский Р.Г., Шепетова О.П. Основы стати-стической оптимизации преподавания иностранных языков. Статистика речи и автоматический анализ текста. Л., 1974. С. 195–234. [Alekseev P.M., German-Prozorova L.P., Piotrovskii R.G., Shepetova O.P. Osnovy statisticheskoy optimizatsii prepodavaniya inostrannykh yazykov (Basics of the Statistical Optimization of Foreign Languages Teaching). Statistika rechi i avtomaticheskiy analiz teksta (Statistics of Speech and Automatic Analysis of the Text). Leningrad, 1974,pp. 195–234.], Salton G. On the Specification of Term Values in Automatic Indexing. Journal of Documentation. 1973, vol. 29, no. 4, pp. 351–372].

Часто именно уникальные термины более точно сигнализируют о теме документа, например, о новизне изобретения в патентных документах.

Для учета параметров частотности и уникальности лексем текста, для вычисления релевантности ключевых слов документа широко использует-ся метод TF-IDF [Jones K.S. A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation. 2004, vol. 60, no. 5, pp. 493–502, Salton G.A Vector Space Model for Automatic Indexing. Communications of the ACM. 1975,vol. 18, no. 11, pp. 613–620.] с применением корпуса одинаковых по тематике документов. Релевантность ключевых слов в данном случае

определяется как произведение двух мер: частоты слова в документе (TF = Term Frequency) и обратной частоты слова в коллекции документов (IDF = Inverse Document Frequency). Последнее означает количе-ство документов в корпусе, где термин употреблен по крайней мере один раз.

Использование корпуса текстов для повыше-ния корректности извлечения ключевых слов по-лучило достаточно широкое распространение, од-нако отсутствие таких корпусов для каждой кон-кретной предметной области в реальной жизни делает применение таких корпусных моделей и методов весьма проблематичным

TF-IDF - статистическая мера, используемая для оценки важности слова, как в контексте документа (TF), так и в контексте корпуса документов (IDF) [14]. Метрика используется с различными вариантами вычисления TF и IDF. Известны различные расширения модели TFIDF, например Okapi BM25. Существует огромное количество исследований и разработок в этой области. . Но следует отметить важную особенность использования TF-IDF - набор данных не должен меняться во время расчета. Это усложняет вычисления, если требуется провести обсчет данных в реальном времени. Для того чтобы решить эту проблему, достаточно актуальную в анализе потоков текстовой информации, например, в социальных сетях, некоторые исследователи предлагают оригинальные модификации классической меры [16], используя богатый опыт существующих разработок. Методы, основанные на применении статистических мер для оценки важности слова в контексте документа или корпуса, можно определить как статистические методы оценки важности слова.

2.1.2.1.2 Методы выделения коллокаций

За последние годы появилось большое число исследований и разработок, посвященных коллокациям, затрагивающих как теоретические аспекты статистического подхода к данному понятию, так и практические методы выявления коллокаций. В этой связи стоит отметить работы [7] и [21], где предлагается описание существующих статистических методов выявления коллокаций, т. е. словосочетаний, обладающих некоторой степенью устойчивости. Самым простым способом выявления коллокаций в тексте является составление частотных списков слов, словоформ и частот совместной встречаемости.

Аппаратом для установления связи между случайной и обусловленной встречаемостью слов служат меры статистической ассоциации: MI, t-score, Log-Likelihood, z-score и др, которые вычисляют силу связи между элементами в составе коллокации и используются в компьютерной лингвистике для выделения ЗС.

Мера MI вычисляет вероятность двух встречающихся вместе слов путем сравнивания произведения их относительных частот в корпусе с наблюдаемыми частотами их совместной встречаемости. Разница между этими величинами выявляет степень значимости их встречаемости.

Если значение MI (n, c)больше единицы, тогда данное сочетание слов является статистически значимым. Если MI (n,c) примерно равнонулю, слова появляются в паре крайне редко. Если MI (n,c) меньше нуля, то n и c находятся в отношении дополнительной дистрибуции. Вопрос о том, какие значения MI считать пороговыми., остается открытым.

Мера MI (mutual information), введенная в работе [Church K., Hanks P. Word association norms, mutual information, and lexicography // Computational Linguistics, 1996. Vol. 16. No. 1. P. 22–29], сравнивает зависимые контекстно-связанные частоты с независимыми частотами

слов в тексте. Если значение MI превосходит определенное пороговое значение, то словосочетание считают статистически значимым. Мера MI вычисляется по формуле:

$$\mathrm{MI} = \log_2 \frac{f(n,c) \times N}{f(n) \times f)c)},$$

где n – первое слово словосочетания; c- второе слово словосочетания; f(n,c) – частота совместной встречаемости двух слов; f(n), f(c) – абсолютные частоты встречаемости каждого слова по отдельности; N - общее число словоупотреблений в корпусе.

Мера t-score также используется при ответе на вопрос, насколько неслучайным является сочетание двух или более слов в тексте. Для вычисления t-score применяется формула:

$$\text{t-score} = \frac{f(n,c) - f(n) \times f(c)/N}{\sqrt{f(n,c)}}.$$

Весьма часто применяется также мера, называемая логарифмическая функция правдоподобия или log-likelihood, выведенная в работе [Dunning T. Accurate methods for the statistics of surprise and coincidence // Computational Linguistics, 1993. Vol. 19. No. 1. P. 61–74.]

Для ее вычисления применяется формула:

$$\text{log-likelihood} = 2 \sum f(n,c) \times \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}.$$

Чаще же всего для выделения коллокаций, особенно терминологических, применяется мера MI [1].[новикова] Основным недостатком этой меры является тот факт, что она является чувствительной к величине корпуса и завышает значимость редких словосочетаний, что приводит к тому, что ее значение будет велико в случае опечаток, иностранных слов и другого

информационного шума, который неизбежен в большой коллекции. Один из вариантов нивелирования данного недостатка - использование порога по частоте.

Основная модификация методов, основанных на статическом подходе, - предварительное использование морфологических шаблонов фильтров [2],[3]

ТакТакого рода модификации могут быть отнесены к гибридным методам.

Пример использования шаблона:

Шаблон	Пример
[прил+сущ.+сущ.(Р.п.)+сущ.(Р.п.)]	*ЦИФРОВАЯ МОДЕЛЬ РЕЛЬЕФА ДНА*
[сущ.+прил.(Р.п.)+сущ.(Р.п.)]	*ИССЛЕДОВАНИЕ МОРСКОГО ДНА*
[прил.+прил.+сущ.]	*БОЕВЫЕ ПОДВОДНЫЕ РОБОТЫ*
С[прил.+сущ.]	*ПОДВОДНАЯ ЛОДКА*
С[прич.+сущ.]	*ОБУЧЕННЫЕ ДЕЛЬФИНЫ*
у[сущ.+сущ.(Р.п.)]	*КАРТА ДНА*
С[сущ.+сущ.(Т.п.)]	*ОБУЧЕНИЕ ДЕЛЬФИНОВ*
у[сущ.+’-‘+сущ.]	*АППАРАТ-РОБОТ*

ществуют и другие статистические подходы для выделения терминологических сочетаний. Например, один из методов заключается в нахождение n-словных сочетаний по заданным частотным характеристикам. Это могут быть значения абсолютных или относительных частот для данных сочетаний слов или значение некоторой статистической меры, согласно которой данная конструкция

была найдена и выдана среди результатов. Далее может быть использован порог отсечения по заданному значению [8].Вопросы информационной значимости списка ключевых слов обсуждаются в [9]. На этом понятии основан так называемый метод координатного индексирования, предполагающий, что основное содержание документа может быть с достаточной степенью точности и полноты выражено соответствующими списками ключевых слов, различающимися частотой употребления. Для определения частоты используется понятие «плотность ключевых слов», выраженное в процентах.

В [5] рассматриваются пять методов автоматического выделения терминоподобных конструкций произвольной длины. Свой подход авторы называют подходом «чистой доски»: на этапе выделения терминов-кандидатов используется минимум информации о структуре и составе терминов, не используются словари, тезаурусы и другие семантические ресурсы, не делаются привязки к определенной предметной области. Приведены результаты автоматической оценки методов с учетом частоты встречаемости кандидатов в термины. Также в этой работе исследовались методы, которые могут быть использованы для выделения терминоподобных словосочетаний произвольной длины и структуры: MaxLen, C-value, k-factor, Window, Синтаксический анализ (АОТ). На основе анализа результатов делается вывод, что сравниваемые методы дают похожие результаты, но использование семантических словарей сочетаемости, продуктивных и непродуктивных слов может существенно повысить качество выделения и сборки терминов.

В заключении стоит сказать, что статистические методы весьма актуальны при изучении новых предметных областей и их терминологии.

2.1.2.2 Лингвистические методы

Попытки создания универсального лингвистического метода извлечения ключевых слов не имели успеха. Различные методы и приемы можно классифицировать по определенным лингвистическим направлениям.

Лингвистические методы основываются на значениях слов, используют онтологии и семантические данные о слове. К сожалению, эти методы слишком трудоемки на ранних этапах: разработка онтологий является очень затратным процессом. К тому же, операции лингвистического анализа текстов, выполняемые вручную, являются источником значительного количества ошибок и неточностей и делают сам процесс анализа документов сложным и длительным. Поэтому необходимым условием является наличие доступных программных средств, позволяющих автоматизировать процесс анализа текстов документов [4].

Примером лингвистических методов извлечения ключевых слов является метод, описанный в [4]. В этой статье предложена методика анализа текстов документов, являющаяся основой автоматизированной информационной системы аудита нормативных документов орган изации, включающая два этапа: исследовательский и аналитический. На исследовательском этапе происходит выделение ключевых слов из текста с использованием словарей (словарь неинформативных лексических единиц, словарь устойчивых словосочетаний и др.) и операций над множествами. Полученные ключевые слова являются исходными данными для второго (аналитического) этапа, целью которого является формулирование рекомендаций по оптимизации бизнес-процессов и электронных документопотоков.

В [6] предлагается метод автоматического извлечения ключевых терминов из текстовых документов, основанный на мере семантической близости терминов, вычисленной с использованием базы данных Википедии, построении семантического графа, выборе тематических терминов при помощи алгоритма Гирвана-Ньюмана. Одним из преимуществ этого метода, по мнению авторов, является отсутствие необходимости в предварительном обучении, так как работает непосредственно с базой данных. Экспериментальные оценки эффективности метода показывают высокую точность и полноту извлечения из текста ключевых терминов.

Лингвистические методы, основанные на применении использования значений слов, словарей, онтологий, энциклопедий, в том числе, Википедии, можно выделить в отдельный метод на основе баз данных и значений слов.

Графовые лингвистические методы представляют большой интерес в области обработки естественного языка, благодаря своей универсальности и эффективности основанных на них алгоритмов [6], [22]. В этих методах основной процедурой является построение семантического графа. Это взвешенный граф, вершинами которого являются термины документа, наличие ребра между двумя вершинами свидетельствует о том, что термины семантически связаны между собой, вес ребра является численным значением семантической близости двух терминов. Ключевые слова отбираются алгоритмами обработки графа. Графовые методы различаются между собой способами отбора множества терминов и определения близости отдельных терминов, которые основаны на статистических параметрах, а также на морфологическом, синтаксическом или семантическом анализе. Вообще

говоря, графовые методы вбирают в себя множество подходов, поэтому некоторые считают , что их можно отнести к гибридным [22].

В [10] используются лингвистические методы, основанные на маркемном анализе. Такой анализ определяет, в каких отношениях находятся интуитивно выделяемые ключевые слова и маркемы. Это позволяет исследовать возможность автоматического выделения ключевых слов. Предлагаемая методика использует доступное программное обеспечение. Делаются выводы, что предложенный алгоритм маркемного анализа научных текстов вполне способен давать достоверную информацию о семантике этих текстов.

В [20] описываются результаты экспериментального исследования процедур автоматического выявления терминов в текстах, основанные на лексико-синтаксических шаблонах языка LSPL.Такой шаблон задает последовательность элементов-слов, из которых должна состоять языковая конструкция, и указывает условия синтаксического согласования этих элементов. Разработаны процедуры выявления терминоупотреблений на основе набора лексико-синтаксических шаблонов. Предложенная стратегия совместного применения этих процедур, позволит повысить F-меру полноты и точности распознавания.

Семантическая близость терминов может определяться разными способами: для английского языка при помощи семантической сети WordNet, для других языков – при помощи аналогичных ресурсов, например, в частности, для русского языка – при помощи RussNet, РуТез [Добров Б.В., Лукашевич Н.В. Тезаурус РуТез как ресурс для решения задач информационного поиска, 2009. http://math.nsc.ru/conference/zont09/reports/93Dobrov-Lukashevich.pdf] и др. новикова

Один из методов данной группы [Гринева М., Гринев М. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов, 2009. http://citforum.ru/database/articles/kw_extraction/] состоит из пяти шагов.

На первом этапе происходит извлечение всех терминов документа и подготовка для каждого термина набора статей какого-либо тезауруса или, например, Википедии, как крупномасштабной и постоянно обновляемой миллионами людей энциклопедии, покрывающей много специфических областей знаний. Также на этом шаге строятся различные морфологические варианты для каждого термина, что позволяет расширить границы поиска по статьям.

На втором этапе для каждого термина выбирается наиболее подходящая статья из всех найденных на предыдущем шаге. Это задача решается, например, при помощи определения контекста слова. Результатом работы данного шага является список терминов, в котором каждый соотнесен с одной статьей.

На третьем шаге строится семантический граф - взвешенный граф, вершинами которого являются термины документа, наличие ребра между двумя вершинами свидетельствует о том, что термины семантически связаны между собой, вес ребра является численным значением семантической близости двух терминов. При этом термины-ошибки, возникшие при разрешении лексической многозначности, оказываются периферийными или даже изолированными.

На четвертом шаге происходит обнаружение сообществ в построенном графе. Это осуществляется, например, при помощи алгоритма Гирвана-Ньюмана, разбивающего граф на подграфы. Для оценки разбиения используется мера модулярности графа, которая является мерой того,

насколько разбиение качественно, т.е. существует много ребер внутри сообщества и мало вне его.

На пятом шаге выбираются те сообщества, которые содержат ключевые термины. Ранжирование основано на использовании плотности и информативности сообщества, где плотность определяется суммой весов ребер сообщества, а информативность – суммой TF-IDF терминов сообщества, деленной на количество терминов сообщества.

2.1.2.3 Гибридные или лингво-статистические методы

Каждый из рассмотренных статистических или лингвистических методов для выделения ключевых слов или значимых словосочетаний (коллокаций) имеет те или иные недостатки. Авторы большинства работ считают, что лишь при сочетании этих методов можно достигнуть наибольшей точности извлечения. Таким образом, при использовании гибридных методик статистические методы обработки документов дополняются одной или несколькими лингвистическими процедурами (морфологическим, синтаксическим, и семантическим анализами) и лингвистическими базами знаний различной глубины (словарями, онтологиями, грамматиками, лингвистическими правилами и т. д.).

Заметим, что как гибридные методы извлечения ключевых слов из документа, так и статистические, могут требовать или не требовать корпуса текстов.

Согласно [22] «большим потенциалом обладают гибридные методики, в которых статистические методы обработки документов

дополняются одной или несколькими лингвистическими процедурами и лингвистическими базами знаний различной глубины».

В работе [12] описываются методы автоматического извлечения двухсловных терминов из отдельного текста или корпуса текстов на основе статистики встречаемости и морфологических шаблонов. Данные методы также используются в технологии Keywen. Показано, что использование совокупности признаков словосочетаний значительно улучшает извлечение терминов.

На основе сравнения существующих систем авторы статьи [13] заключают, что для улучшения извлечения терминов и получения более релевантных результатов (уменьшения информационного шума), должны быть выполнены следующие условия: проведены лингвистически ориентированные исследования семантических связей терминов и условий ограничения терминологических единиц в пределах данной специальной области и в данном текстовом типе; программные системы должны научиться сочетать статистические и лингвистические методы и поддерживать более одной стратегии. Также должна быть полезной разработка общей шкалы тестирования и оценки/сравнения качества извлекаемых терминов.

Статья [8] представляет результаты исследования по выделению терминологических словосочетаний на основе статистических мер и грамматики синтаксических конструкций с помощью специализированной системы обработки корпусных данных. Описываются эксперименты по автоматическому выделению терминов и терминологических сочетаний. Оценивается эффективность разных подходов. Фактически этот подход является комбинированным, так как объединяет лингвистический и статистический методы.

Результаты разнообразных исследований в области гибридных лингво-статистических методов см. также в [15].

К числу гибридных методов извлечения ключевых слов можно отнести методы на основе машинного обучения, где задача извлечения ключевых слов рассматривается как задача классифика-ции. Методы на основе машинного обучения для создания обучающей выборки и построения модели-классификатора, как правило, требуют корпуса документов с размеченными ключевыми словами.

Помеченные ключевые слова считаются положи-ельным примером, остальные слова – отрицательным примером. Далее высчитывается релевантность каждого слова тренировочного текста путем сопоставления ему вектора значений различных параметров, например, меры TF-IDF, длины слова, части речи, положения слова в заголовке, положения слова в первом абзаце, последнем абзаце, в списках литературы и т. д. Фиксируются отличие значений векторов этих параметров для ключевых слов и не ключевых. Далее вычисляется вероятность отнесения каждого слова к группе ключевых и задается ее порог, т. е. модель обучается. Извлечение ключевых слов из нового документа происходит путем вычисления релевантности слов и их вероятности отнесения к ключевым в соответствии с построенной моделью.

Среди методов на основе машинного обучения можно отметить:

- байесовские методы [8, 18, 41, 38];
- метод опорных векторов [13, 15, 19];
- деревья решений [37];
- использование нейронных сетей [22, 33, 40].

Анализ существующих методов автоматиче-ского извлечения ключевых слов показывает, чтодля создания эффективных экстракторов ключевых

слов следует учитывать, лингвистические типы естественных языков (аналитический, флектив-ный, агглютинативный, изолирующий), предметную область (подъязык) и наличие необходимых лингвистических и программных ресурсов.

2.1.2.3.1 Метод дистрибутивной семантики

Метод дистрибутивной семантики тоже относится к гибридным методам. Модели дистрибутивной семантики нашли применение в исследованиях и практических реализациях, связанных с семантическими моделями естественного языка.

Дистрибутивные модели применяются для решения следующих задач:

- выявление семантической близости слов и словосочетаний [Sketch Engine corpus manager (англ.). Lexical Computing Ltd.. Проверено 17 апреля 2016.];
- автоматическая кластеризация слов по степени их семантической близости;
- автоматическая генерация тезаурусов и двуязычных словарей[15][17];
- разрешение лексической неоднозначности;
- расширение запросов за счет ассоциативных связей;
- определение тематики документа;
- кластеризация документов для информационного поиска;
- извлечение знаний из текстов;
- построение семантических карт различных предметных областей[7];
- моделирование перифраз;
- определение тональности высказывания;

- моделирование сочетаемостных ограничений слов [18

Дистрибутивная семантика занимается вычислением степени семантической близости между лингвистическими единицами на основании их дистрибуционных признаков в больших массивах лингвистических данных. Дистрибутивный метод предполагает учет отношений между языковыми единицами, их распределение в тексте. Метод считается достаточно новым и относится к одному из наиболее современных приемов лингвистического анализа, так как стал широко применяться при "исследовании семантических явлений только в конце 20 века.

Основные сферы применения дистрибутивных моделей: разрешение лексической неоднозначности, информационный поиск, кластеризация документов, автоматическое формирование словарей (словарей семантических отношений, двуязычных словарей), создание семантических (визуальных) карт, определение тематики документа и др.

Как правило, значимые единицы языка распределяются в речи не произвольно, а по определенным закономерностям, которые могут иметь детерминистический (логический, точно предсказуемый) или, что гораздо чаще, вероятностный характер. В основе обоих видов закономерностей распределения лежат семантические свойства языковых единиц. В свою очередь, эти семантические свойства подразделяются на общие грамматические и индивидуальные смысловые. В то время как грамматическая сочетаемость подчиняется преимущественно детерминистическим правилам, лексическая сочетаемость определяется главным образом вероятностными правилами. Такой вид распределения слов в связном тексте, как совместная встречаемость, фактически полностью подчиняется вероятностным закономерностям.

Дистрибутивно-трансформационный анализ позволяет расклассифицировать синтаксически активные слова на семантические группы, но фактически ничего не говорит о строении (структуре) этих групп, характере семантических связей между членами группы. К тому же получаемые классы слов не представляются единственно возможными со смысловой точки зрения, поскольку лексические единицы могут объединяться в семантические группы на других основаниях, например на предметно-логических (тематические группы), морфологических (словообразовательные гнезда), формально-грамматических и др. Причем получаемые на разных основаниях таксономии не обязательно должны совпадать между собой.

Для определения силы семантической связи между членами заданной лексической подсистемы достаточно эффективной представляется дистрибутивно-статистическая методика А. Е. Супруна. Суть этой методики заключается в статистическом анализе слов, сочетающихся с рассматриваемыми лексемами в текстах. В лексической сочетаемости проявляются семантические свойства слов, поэтому, чем больше элементов лексической сочетаемости учитывается для каждого исследуемого слова, тем полнее представляются его семантические свойства. В этом случае задача заключается в том, чтобы, изучая тексты достаточно большого и статистически достоверного объема, фиксировать в них для заданных слов все их семантические свойства, выявляемые через лексическую сочетаемость. Получаемые при этом данные предоставляют возможность составить для рассматриваемых слов своеобразную семантическую анкету, в которой дается описание значений слов путем набора реализованных ими в текстах семантических признаков. Вес или значимость каждого семантического признака измеряется частотой слов, которые сочетаются с исследуемым и

выражают этот признак (например, признак размера выражается прилагательными *большой, огромный, громадный, маленький* и т. д., признак движения - глаголами *бежать, ехать, мчаться, идти* и т.п.). Сравнивая значения заданных слов по совокупности выявленных таким образом семантических признаков, исследователь может с помощью статистических формул измерять степень сходства семантики слов.

Сильной стороной этой разновидности дистрибутивно-статистического анализа служит то, что исследование строится на большом фактическом материале, позволяющем измерять силу семантических связей между словами и на этом основании вскрывать структуру лексических подсистем, решать некоторые проблемы синонимии и антонимии (определять меру синонимичности и антонимичности интересующих нас слов), измерять семантический объем слов и т. п. К факторам, ограничивающим сферу применения этой методики, относится то, что процесс исследования достаточно трудоемок, методика пригодна только для регулярно употребляемых в текстах слов.

Следующая разновидность дистрибутивного метода направлена на учет совместной встречаемости лексических единиц в заданном интервале текста (чаще всего в пределах от одноместного окружения исследуемой единицы до предложения или абзаца). Фактически в основе этой методики лежит идея, сформулированная Ф. Ф. Фортунатовым: "Чем чаще сочетаются в опыте известные духовные явления или чем сильнее они в этом сочетании, тем больше они способны воспроизводить впоследствии одно другое, и наоборот, чем реже они сочетаются в опыте или чем слабее духовные явления в этом сочетании, тем менее способны воспроизводить они впоследствии одно другое" (Фортунатов 1956: 113). На материале письменных текстов эта идея реализуется в виде совместной встречаемости тех слов, которые связаны между собой по

значению, т. е. *чем сильнее семантическая связь между словами, тем чаще встречаются они в тексте недалеко друг от друга*. Сила семантической связи измеряется в этом случае отклонением конкретной зафиксированной в текстах частоты совместной встречаемости слов от теоретически ожидаемой. Эксперименты, проведенные на материале английских текстов с помощью рассматриваемой методики, продемонстрировали ее эффективность как для выделения семантических полей, так и для вскрытия их структуры (Шайкевич 1963).

Методика, направленная на учет частоты совместной встречаемости слов, может опираться на различные исходные положения: во-первых, можно регистрировать совместиую встречаемость всех слов подряд и самые частые встречаемости затем подвергнуть лингвистическому анализу; во-вторых, допустимо фиксировать совместную встречаемость только тех слов, которые интересуют исследователя, например определенных синонимов, слов, принадлежащих к одной тематической

Каждый из приемов дистрибутивного анализа может быть применен для решения конкретных семантических проблем, поскольку именно через лексическую сочетаемость и совместную встречаемость слов реализуются их разнообразные семантические свойства, т.е. проявляются те элементы значения, которые существенны для функционирования слова и для отношений между словами. Ценность результатов, полученных при использовании различных дистрибутивных методов, заключается в том, что эти результаты выводятся из лингвистической данности, в которой отражены и зафиксированы самые разнообразные аспекты языкового значения.

Употребляемые при исследовании семантических явлений разновидности дистрибутивного метода могут быть представлены в виде схемы (рис. 16).

Рис. 16

Дистрибутивная семантика базируется на гипотезе о том, что лингвистические элементы со схожей дистрибуцией имеют близкие значения [Turney P. D., Pantel P. From frequency to meaning: Vector space models of semantics. // Journal of Artificial Intelligence Research (JAIR), 2010, №37. P. 141-188]. В основе всех современных вариантов дистрибутивного подхода лежат количественные оценки, которые характеризуют совместную встречаемость языковых единиц текста в контекстах определенной величины.

Каждому слову присваивается свой *контекстный вектор*. Множество векторов формирует *словесное векторное пространство*.

Семантическое расстояние между понятиями, выраженными словами естественного языка, вычисляется как расстояние между векторами словесного пространства.

Идея контекстных векторов была предложена психологом Ч. Осгудом в рамках работ по представлению значений слов[Osgood et al., 1957.].

Контексты, в которых встречались слова, выступали в качестве измерений многоразрядных векторов.

Термин контекстный вектор был введён С. Галлантом для описания смысла слов и разрешения лексической неоднозначности [*Sahlgren M.* The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces (Ph.D. Thesis). — Department of Linguistics, Stockholm University, 2006.].

Была разработана дистрибутивно-семантическая методика и соответствующее программное обеспечение, которые позволяют автоматически сравнивать контексты, в которых встречаются изучаемые языковые единицы, и вычислять семантические расстояния между ними[*Sahlgren M.* The Distributional Hypothesis. From context to meaning (англ.) // Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), Rivista di Linguistica : журнал. — 2008. — Vol. 20, no. 1. — P. 33-53.].

2.2.1.6.1. Модель Векторного пространства

В модели используются векторные пространства из линейной алгебры. Семантическое расстояние между понятиями, выраженными словами естественного языка, вычисляется как расстояние между векторами словесного пространства.

Информация о дистрибуции лингвистических единиц представляется в виде многоразрядных векторов, которые образуют словесное векторное пространство, а семантическая близость между лингвистическими единицами вычисляется, как расстояние между векторами.
Многоразрядные вектора образуют матрицу, где каждый вектор соответствует лингвистической единице (слово или словосочетание), а каждое измерение вектора соответствует контексту (документ, параграф, предложение, словосочетание, слово).

Для вычисления меры близости между векто-рами могут использоваться различные формулы:

расстояние Минковского, расстояние Манхеттена, евклидово расстояние, расстояние Чебышёва, скалярное произведение, косинусная мера.

Для вычисления меры близости между векторами наиболее популярна косинусная мера, в которой обычно используются характеристики совместной встречаемости пар слов и одиночной встречаемости каждого из слов.

$$\frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

где A и B — два вектора, расстояние между которыми вычисляется.

После проведения подобного анализа становится возможным выявить наиболее близкие по смыслу слова по отношению к изучаемому слову.

Lemma	Score	Freq
кот	0.3	9252
собака	0.288	24102
птица	0.219	13889
зверь	0.215	7270
пес	0.214	4820
животное	0.21	16108
волк	0.199	6071
мальчик	0.198	28828
девочка	0.197	27136
медведь	0.196	5286
крыса	0.186	4021
парень	0.174	25625
мама	0.172	42197
корова	0.168	5466
папа	0.164	22231
лошадь	0.164	12582
мышь	0.164	4887

Существует множество различных моделей дистрибутивной семантики, которые различаются по следующим параметрам:

- тип контекста: размер контекста, правый или левый контекст, ранжирование;
- количественная оценка частоты встречаемости слова в данном контексте: абсолютная частота, TF-IDF, энтропия, совместная информация и пр.;
- мера расстояния между векторами: косинус, скалярное произведение, расстояние Минковского и пр.;
- метод уменьшения размерности матрицы: случайная проекция, сингулярное разложение, случайное индексирование и пр.

Модель, которая была задействована в настоящей работе, отличается следующими параметрами:

- тип контекста:контекст в виде коллекции предложений;
- абсолютная частота, как количественная оценка встречаемости слова в данном контексте ;
- совместная информация;
- мера расстояния между векторами: (косинус угла)
- метод уменьшения размерности матрицы: сингулярное разложение.

Наиболее широко известны такие дистрибутивно-семантические модели как Модель векторных пространств, Латентно-семантический анализ, Тематическое моделирование и Предсказательные модели

При применении дистрибутивно-семантических моделей в реальных приложениях возникает проблема слишком большой размерности векторов, соответствующей огромному числу контекстов,

представленных в текстовом корпусе. В связи с этим требуется применение специальных методов, которые позволяют уменьшить размерность и разреженность векторного пространства и при этом сохранить как можно больше информации из исходного векторного пространства. Получающие в результате сжатые векторные представления слов в англоязычной терминологии носят название *word embeddings.*

Методы, используемые для уменьшения размерности векторных пространств включают в себя:

- удаление определенных измерений векторов в соответствии с лингвистическими или статистическими критериями;
- сингулярное разложение;
- метод главных компонент (PCA);
- случайное индексирование[11].

В настоящее время существует несколько программных средств для проведения исследований по дистрибутивной семантике с открытым кодом это S-Space, Semantic Vectors, Gensim, word2vec, WebVectors

Модели векторных пространств находят все более широкое применение в исследованиях, связанных с семантическими моделями ЕЯ, и имеют разнообразный спектр приложений [Lenci A. Distributional semantics in linguistic and cognitive research. // Rivista di Linguistica, 2008. Vol. 1. P. 1–30; Charnine M., Charnine V. Keywen category structure. // Wordclay, USA, 2008. 60 p.]. Данная область в настоящее время является одной из наиболее актуальных. Наиболее известная модель дистрибутивной семантики – это латентный семантический анализ, разработанный для решения проблемы синонимии при информационном поиске [5], и модель языка как гиперпространства, разработанная как модель семантической памяти человека [6].

Следует отметить модели Word-Space Model [SahlgrenM. The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. Thesis. // Department of Linguistics, Stockholm University, 2006] и Semantic Space Model [Sahlgren M. Towards pertinent evaluation methodologies for word-space models // LREC 2006: 5th Conference (International) on Language Resources and Evaluation Proceedings. // Genoa, Italy, 2006].

Семантическая информация извлекается из текстов большого объема на основе анализа окружения слова. Слово рассматривается как точка в многомерном семантическом пространстве. На основе близости между точками семантического пространства вычисляется семантическое сходство между словами с использованием метрик. Анализ семантического сходства выполняется на основе статистических методов с расчетом частотности появления в тексте близких точек семантического пространства. По результатам исследования делается вывод о том, что контекстное окружение играет важную роль в распознавании лексических отношений между словами.

Концепция СВП впервые была реализована в информационно-поисковой системе SMART [The SMART retrieval system: Experiments in automatic document processing. / Ed. G.M. Salton. // Prentice-Hall, 1971].

Идея СВП состоит в представлении каждого документа из коллекции текстов в виде точки в многомерном семантическом пространстве, которой соответствует вектор в векторном пространстве значимых словосочетаний. Точки, расположенные ближе друг к другу в этом пространстве, считаются более близкими по смыслу. Пользовательский запрос рассматривается как псевдодокумент и тоже представляется как точка в этом же пространстве. Документы сортируются в порядке возрастания расстояния, т. е. в порядке уменьшения семантической близости от запроса, и в таком виде предоставляются пользователю.

В настоящее время СВП используют для измерения степени близости запроса и найденных документов [Manning C., Raghavan P., Sch.utze H. Introduction to information retrieval. Cambridge: Cambridge University Press, 2008]. Baroni и Lenci [Baroni M., Lenci A. Distributional memory: A general

framework for corpus-based semantics. Comput. Linguistics, 2010. Vol. 36. Iss. 4. P. 673–721] предложили обобщенную модель, названную «дистрибутивная память», которая является обобщением ранее известных моделей векторных пространств (vector spaces), семантических пространств (semantic spaces), пространств слов (word spaces), семантических моделей корпусной статистики (corpus-based semantic models) и дистрибутивных семантических моделей (distributional semantic models).

Впоследствии концепция СВП успешно применялась для других семантических задач. Так , в работе [8] контекстное векторное простран-ство было использовано для оценки семантической близости слов. Данная система достигла результата 92,5% на тесте по выбору наиболее подходящего синонима из стандартного теста английского язы-ка TOEFL, в то время как средний результат при прохождении теста человеком был 64,5%.

В настоящее время ведутся активные исследо-вания по унификации модели СВП и выработке общего подхода к различным задачам выявления семантических связей из корпусов текстов [9].

Модели СВП нашли применение для построения концептуальных моделей предметных областей.

В данной работе СВП используется для нахожления значимых словосочетаний (ЗС) ПО и ассоциативных связей между ними. Под ЗС понимаются лексические последовательности, имеющие тенденцию к совместной встречаемости.

Модель СВП, которая используется в данной работе, обладает следующими признаками:

- Тип изучаемых единиц: значимые словосочетания;
- Тип контекста: лексемы и словосочетания, размер контекста – предложение, ранжирование контекста – нет;

- Количественная оценка частоты встречаемости изучаемой единицы в данном контексте: абсолютная частота;
- Метод вычисления расстояния между векторами: косинусная мера.

Приведем пример построения СВП на основе следующиего фрагмента:

Построим контекстные векторы для ЗС: «компьюторная лингвистика», «искусственный интелект», «дискретная математика», «конструктивная математика» и слов, встречающихся в текстовом фрагменте более одного раза. В таблице используются сокращенные обозначения: с1 - искусственный интелект, с2-компьюторная лингвистика, с3 -дискретная математика, с4-конструктивная математика, с5 –интеллект, интеллектуальный, с6 – математика, математический, с7 – изучать, изучение.

Применив формулу вычисления косинусной меры между контекстными векторами, получаем

	C1	C2	C3	C4	……
C1	0	1	0	0	
C2	1	0	0	0	
C3	0	0	0	0	
C4	0	0	0	0	
…….					

следующие коиффициэнты семантической близости между рассматриваемыми ЗС:

- «дискретная математика» и «конструктивная математика» – 0.97;
- «искусственный интеллект» и «компьюторная лингвистика» – 0.7;
- «компьюторная лингвистика» и «дискретная математика» – 0.52;
- «компьюторная лингвистика» и «конструктивная математика» – 0.4;
- «искусственный интеллект» и «дискретная математика» – 0.36;

- «искусственный интеллект» и «конструктивная математика» – 0.29.

Следует отметить, что при обработке больших массивов лингвистических данных можно создавать различные лингвистические ресурсы, в том числе семантические словари и семантические карты ПО.

2.2.1.6.2 **Тематический анализ**

Семантический анализ больших коллекций неструктурированного текста является актуальной проблемой. Тематическое моделирование — одно из приложений машинного обучения к анализу текстов, которое определяет тематическую структуру каждого документа из текстовой коллекции, а также тематический профиль всех слов этой коллекции. Современный анализ текстов практически всегда работает с большими объемами данных, которые можно обработать лишь с помощью параллельных или распределенных реализаций алгоритмов тематического моделирования.

На данный момент существует много методов семантического анализа корпуса текстов, включая методы кластеризации, методы тематического моделирования LSA, PLSA, PLSI, LDA, hLDA, ITM и др… Все эти методы используются для нахождения сильно связанных КТ. Метод hLDA может быть использован для рассчета иерархической структуры связей между кластерами.

Известно значительное количество работ по автоматическому извлечению семантических связей из больших массивов текстов на ЕЯ. Некоторые виды статистических тематических моделей могут основываться на традиционных методах автоматической кластеризации текстов [Q. He, K. Chang, E. Lim, A. Banerjee. Keep It Smile with Time: A Reeximanation of Probabilistic Topic Detection Models. In the Proceedings of IEEE Transactions Pattern Analysis and Machine Intelligence. Volume 32, Issue 10, pp. 1795–1808, 2010]. Наиболее успешными методами считаются Hierarchical Latent

Dirichlet Allocation (hLDA), модели дистрибутивной семантики и модели семантических векторных пространств (СВП).

В последнее время предложены вероятностные механизмы выделения тем в текстовых коллекциях, например, методы, основанные на скрытом распределении Дирихле (Latent Dirichlet allocation), которые в настоящее время интенсивно исследуются в рамках различных приложений автоматической обработки текстов. [D. Blei, A. Ng and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, No 3, pp. 993–1022, 2003]. D. Blei, A.Y. Ng, and M.I. Jordan (2003) предложили генеративную модель для анализа текстовых и других коллекций дискретных данных, утверждая, что каждый документ генерируется в виде смеси тем, где непрерывные многозначные пропорции смешения распределяются как скрытое распределение случайных величин Дирихле.

На сегодняшний день разработано довольно много тематических моделей. Для выбора наиболее эффективной модели проанализированы некоторые работы, в которых выполняется сравнение моделей с точки зрения практических приложений. Так, в работе K. Stevens и др. (2012) авторы сравнивали между собой методы NMF (неотрицательной матричной факторизации) и LDA (латентного размещения Дирихле) и пришли к выводу, что эти алгоритмы дают похожее качество, хотя NMF выдаёт немного больше бессвязных подтем [K. Stevens, P. Kegelmeyer, D. Andrzejewski, D.Buttler. Exploring Topic Coherence over many models and many topics. In the Proceedings of EMNLP-CoNLL, pp. 952–961, 2012].

Традиционные тематические модели, как правило, основаны на методах жёсткой кластеризации, рассматривающих каждый документ как разреженный вектор в пространстве слов большой размерности [G. Salton. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley, 1989]. Алгоритм NMF, изначально разработанный для уменьшения размерности,

осуществляет нечёткую кластеризацию, которая относит один и тот же документ к разным кластерам с разными вероятностями.

Для обработки неструктурированной текстовой информации можно использовать модель LDA, позволяющую автоматически классифицировать новые текстовые документы, оценивать сходство между документами в целях поиска информации и создавать вероятностную модель большой коллекции текстов.

Тематические модели, иллюстрированные примером скрытого распределения Дирихле, весьма симпатичны, поскольку они обнаруживают группы слов, которые часто появляются вместе в документах. Эти группы являются полиномиальными распределениями слов по лексике. Слова, которые имеют наибольшую вероятность в теме, показывают содержание темы.

LDA — это иерархическая байесовская модель, которая предполагает, что имеются множество из K фиксированных скрытых тем, в соответствии с которыми генерируются документы, и что каждая тема представлена как некоторое распределение |V| слов в словаре. Документ моделируется путем смешения некоторых тем из общего количества, а затем выборки слов из полученной смеси.

На выходе после обучения модели LDA получаются векторы θ, показывающие, как распределены темы в каждом документе, и распределения β, демонстрирующие, какие слова более вероятны в тех или иных темах. Таким образом, из результатов LDA легко получить для каждого документа список встречающихся в нём тем, а для каждой темы – список характерных для неё слов, т.е. фактически описание темы. Кроме того, можно оценить, какие слова наиболее характерны для той или иной темы – то есть выделить группу слов, максимально релевантных для соответствующей темы.

В настоящей работе для улучшения параметров иерархической кластеризации терминов используются модели LDA совместно с моделями дистрибутивной семантики и семантических векторных пространств

Иерархическая кластеризация - совокупность алгоритмов упорядочивания данных, основывающихся на том, что некоторое множество объектов характеризуется определенной степенью связанности. Предполагается наличие вложенных групп (кластеров различного порядка). В данном классе алгоритмов обработки данных мы использовали алгоритмы тематического моделирования. Их задача - построение модели коллекции документов, которая позволит определить темы, к которым относится каждый из документов. Основные сферы применения тематического моделирования: тематический поиск, классификация, суммаризация и аннотация коллекций документов и новостных потоков. Первое описание тематического моделирования появилось в работе Рагавана, Пападимитриу, Томаки и Вемполы в 1998 году. Томас Хофманн в 1999 году предложил вероятностное скрытое семантическое индексирование (PLSI).

Одна из самых распространенных тематических моделей — это латентное размещение Дирихле (LDA), эта модель является обобщением вероятностного семантического индексирования. Модель иерархического скрытого размещения Дирихле (Hierarchical Latent Dirichlet Allocation, hLDA), описанная в работе [2], базируется на вложенном процессе китайского ресторана (Nested Chinese Restaurant Process, nCRP).

Другие тематические модели, как правило, являются расширением LDA, например, размещение Патинко улучшает LDA за счёт введения дополнительных корреляционных коэффициентов для каждого слова, которое составляет тему.

Одним из недостатков LDA является тенденция к извлечению очень общих тем для данного набора документов. В том случае, когда наряду с основной концепцией присутствует некоторое количество сопутствующих ей неосновных смысловых компонентов в результатах работы LDA с большой вероятностью будет находиться тема, которая включает как основной смысл концепции, так и все сопутствующие ей компоненты, что нередко требуется для решения некоторых задач. В таких случаях употребляются иерархические направленные на определенную тематику модели, дозволяющие моделировать иерархию тем — от более общих до самых узких. Иерархический процесс Дирихле можно понимать как процесс Дирихле над процессом Дирихле. Иначе говоря, при генерации случайного элемента с помощью иерархического процесса Дирихле сначала решается вопрос, из какого процесса Дирихле верхнего уровня следует генерировать элемент, после чего к выбранному процессу Дирихле применяется стандартная схема генерации элемента. Он моделирует документы, в которых имеются как общие для всего набора данных темы, так и более узкие (специфичные).

2.1.3.3. Спектральный анализ

В общем случае, спектр документа представляет собой множество пар (базовая форма слова или словосочетания, число вхождений этого слова или словосочетания в текст). Спектрограмма текста может быть легко построена с применением несложного программного обеспечения. Дальнейшая работа по выделению ключевых слов строится на основе анализа спектрограмм различных текстов, относящихся к предметной области. При этом понятие "словоформа" не используется в алгоритмах анализа.

Оригинальный метод поиска ключевых слов в текстовых файлах, основанный на спектральном анализе, предлагается в работе [11]. Подход базируется на использовании преобразований Фурье и так называемых «карт усреднения» - подобий частотно-временного представления сигнала, полученного с помощью вейвлет-преобразования. Предлагаемый метод работает со спектрограммой текста, в которой содержится в новом качественном виде содержимое текстового файла. В работе описан спектральный алгоритм поиска ключевых слов. Автор обращает внимание на то, что этот метод лишен многих недостатков, присущих другим методам быстрого поиска ключевых слов в текстовых документах. Например, не требуется построений сложных индексов и словарей, использующих B+, суффиксные или GiST деревьяпоиска и требующих довольно больших аппаратно-временных ресурсов. Также не требуется применять сложные процедуры лексического и морфологического анализа. Кроме того, метод позволяет определить местоположение ключевого слова в тексте (классическая модель векторного пространства документов не дает такой возможности).

3 Результаты эксперемента

Результаты эксперемента ПО АНПА

В методе используется понятие «семантическое контекстное пространство», выраженное термином, где точки пространства соответствуют контекстным векторам не отдельных терминов, а значимых словосочетаний (ЗС) (понятие более широкое, чем СВП). Такое семантическое контекстное пространство, представляющее множество связей между ЗС в некоторой предметной области называется ассоциативно-иерархическим портретом предметной области (АИППО).

Значимые словосочетания - это лексические последовательности, имеющие тенденцию к совместной встречаемости (в лингвистике используется термин «коллокация»). Для выделения ЗС в компьютерной лингвистике используются различные статистические меры ассоциативной связанности (association measures), вычисляющие силу связи между элементами в составе коллокации.

По сравнению с существующими методиками, рассматриваемая модель из изначального семантического контекстного пространства, не связанного с конкретной тематикой, автоматически выбирает ту или иную предметную область и её компоненты: значимые словосочетаний (ЗС), ассоциативные связи и, соответственно, контексты для их выделения. В результате строится система множественных ассоциативных связей, а затем формируется ассоциативно-иерархический портрет предметной области – АИППО.

Разработчики СВП отмечают, что основная проблема метода заключается в трудностях учета порядка слов, составляющих контексты. В рамках данного проекта эта проблема решается путем перехода от контекста слов к контексту значимых словосочетаний.

Для построения иерархических связй используется метод LDA, который позволяет объединять близкте по смыслу термины в группы, называемые темами. Также испольщуется метод hLDA для построения иерархии тем. В методе LDA впервые были построены темы, состоящие не только из набора слов, но содержащие также значимые словосочетания (ЗС). ЗС позволяют учесть порядок слов, что является существенным улучшением метода LDA, основанного на идеологии “мешка слов”.

Исследование иерархических связей базируется на гипотезе о том, что более общие термины, занимающие в иерархии категорий более высокое место, имеют больше ассоциативных связей и более высокую частоту

встречаемости будет вестись с помощью поиска соответствующих лексико-синтаксических шаблонов

Для автоматического построения АИППО предлагаются методики автоматического выявления значимых терминов (ЗТ), которые в общем случае рассматриваются как лексические последовательности, имеющие самостоятельную значимость, и которые определяются по абсолютной частоте их встречаемости в текстах ПО.

Для выделения ЗТ используются методы статистического ранжирования. Предполагается также выявление ассоциативных и иерархических связей между ЗТ ПО. Для расчета силы ассоциативной связи ЗТ используется косинусная мера между контекстными векторами (компонентами вектора ЗТ являются частоты совместной встречаемости данного ЗТ с другими ЗТ в одном и том же контексте). Следует отметить, что посредством ЗТ могут быть представлены как отдельные слова, словосочетания и термины, так и более сложные конструкции – объекты и именованные сущности.

Иерархические связи между ЗТ выбираются из числа ассоциативных связей таким образом, чтобы более общие ЗТ имели большее количество ассоциативных связей, при этом учитываются лексико-синтаксические шаблоны, объединяющие данные ЗТ.

Для создания дайджеста текстов – обучающей выборки по ПО «АНПА» был использован поисковый запрос, состоящий из первоначального набора ключевых слов. Запрос включал в себя набор терминов из двух составляющих: 1) субъекты исследования, 2) объекты исследования.

Итогом работы системы семантического серфинга явился расширенный список ключевых фраз, а также дайджест, состоящий из предложений (ключевых фраз), удовлетворяющих поисковому запросу, объёмом около 40000, составленный из такого рода ключевых фраз:

- Конструкция АНПА Remus 100 позволяет устанавливать на него различные сонары, сенсоры и камеры, которые могут быть использованы при картографировании рельефа морского дна, для научных исследований, обнаружения мин и мусора.
- Проект крупнотоннажного АНПА получил название LDUUV (Large Displacement Unmanned Underwater Vehicle) и будет реализовываться в два этапа общей продолжительностью свыше 4,5 года.
- Ранее ВМС Великобритании использовали для обнаружения мин и картографирования морского дна АНПА Remus.
- Японские военные намерены использовать подводных роботов для составления карты дна, изменившегося после прошлогоднего землетрясения, а также в целях противоминной борьбы.
- Этот эхолот обладает двухлучевым излучателем, что позволяет видеть большую поверхность дна, а значит находить большее количество рыбы.
- В конце 2011 года в дополнение к британским АНПА Черноморский флот закупил телеуправляемый автономный необитаемый подводный аппарат "Обзор-600" производства российской компании "Тетис-ПРО".
- Военно-морские силы Великобритании приняли на вооружение автономный необитаемый подводный аппарат (АНПА) Recce, который будет использоваться для обнаружения мин.
- Спасательные силы Черноморского флота России приняли на вооружение новый телеуправляемый автономный необитаемый подводный аппарат "Обзор-600", созданный российской компанией "Тетис-ПРО".
- Как сообщает Aviation Week, программа под названием LDUUV (Large Displacement Unmanned Underwater Vehicle) будет разделена два этапа, общей продолжительностью

В целом было выделено 2298 ключевых фраз, содержащих ключевые термины, которые в совокупности образовали корпус текстов ПО «АНПА».

Эта БД имеет значительный объем – до нескольких Тб. Следует отметить , что для ее хранения также необходимо применять специальные программно-аппаратные технологии big data.

По полученной БД производятся статистические подсчеты с целью выделения ключевых слов, значимых словосочетаний и определения их иерархических связей.).

Фрагмент ключевых слов можно увидеть ниже:

ПОДВОДНЫЕ РОБОТЫ	*DEEP FLIGHT SUPER FALCON*
ПОДВОДНЫЙ РОБОТ	*ODYSSEY MARINE EXPLORATION*
АНПА	*БАССЕЙНОВЫХ ИСПЫТАНИЙ*
ПОДВОДНЫЙ АППАРАТ	*БАССЕЙНОВЫЕ ИСПЫТАНИЯ*
МОРСКОЕ ДНО	*ГИДРОАКУСТИЧЕСКОЕ ОБСЛЕДОВАНИЕ*
ОБСЛЕДОВАТЕЛЬСКИЕ ЗАДАЧИ	*ФОТОТЕЛЕВИЗИОННОЕ ОБСЛЕДОВАНИЕ*
ГИДРОЛОКАТОР	*ДОННАЯ ПОВЕРХНОСТЬ*
ПОВЕРХНОСТЬ ДНА	*ДОННОЙ ПОВЕРХНОСТИ*
ПОВЕРХНОСТИ ДНА	*ПОДВОДНЫЕ ИССЛЕДОВАНИЯ*
МОДЕЛЬ РЕЛЬЕФА	*СТАЦИОНАРНОЕ ТЕЧЕНИЕ*
ОБРАБОТКА ИЗОБРАЖЕНИЙ	*СТАЦИОНАРНОГО ТЕЧЕНИЯ*
GAVIA	*МНОГОЛУЧЕВОЙ ЭХОЛОТ*
TELEDYNE GAVIA	*ЭХОЛОКАЦИОННАЯ СИСТЕМА*
КЛАВЕСИН	*ДОПЛЕРОВСКИЙ ЛАГ*
ИПМТ	*УКБ ГАНС*
BLUEFIN ROBOTICS	*ГРАББЕР*
OCEANSERVER TECHNOLOGY	*МЕЛКОВОДНЫЕ ИСПЫТАНИЯ*
ЗАТОНУВШИЕ ОБЪЕКТЫ	*ПРЕДЕЛЬНАЯ РАБОЧАЯ ГЛУБИНА*
ЗАТОНУВШИХ ОБЪЕКТОВ	*ПРЕДЕЛЬНУЮ РАБОЧУЮ ГЛУБИНУ*
ЧЕРНЫЕ ЯЩИКИ	*ГИДРОДИНАМИЧЕСКИЙ РАСЧЕТ*
ИГП	*ГИДРОДИНАМИЧЕСКОГО РАСЧЕТА*
ГИДРОЛОГИЧЕСКИЕ ПАРАМЕТРЫ	*ПОДВОДНО-ТЕХНИЧЕСКИХ СООРУЖЕНИЙ*
ЭХОЛОТ	*ИНТЕРНЭШНЛ САБМАРИН ИНДЖИНИРИНГ*
ЦМР	*БОЕВЫЕ ПОДВОДНЫЕ РОБОТЫ*
ПОДВОДНАЯ РОБОТОТЕХНИКА	*МОРСКОМ ДНЕ*

ПОДВОДНОЙ РОБОТОТЕХНИКИ	*UNMANNED UNDERWATER VEHICLE*
ПОДВОДНЫЕ РОБОТЫ	*ВОДОЛАЗНОЕ ОБОРУДОВАНИЕ*
НЕОБИТАЕМЫЕ ПОДВОДНЫЕ АППАРАТЫ	*ЗАТОНУВШИХ КОРАБЛЕЙ*
НЕОБИТАЕМЫЙ ПОДВОДНЫЙ АППАРАТ	*ЗАТОНУВШИЕ КОРАБЛИ*
НЕОБИТАЕМОГО ПОДВОДНОГО АППАРАТА	*ДВУХЛУЧЕВОЙ ЭХОЛОТ*
МОРСКОГО ДНА	*ДВУХЧАСТОТНЫЙ ЭХОЛОТ*
ПОДВОДНЫЙ ТЕЛЕУПРАВЛЯЕМЫЙ АППАРАТ	*ГЛУБОКОВОДНОГО ПОГРУЖЕНИЯ*
ПТА	*ГЛУБОКОВОДНОЕ ПОГРУЖЕНИЕ*
СУПЕРГНОМ	*ПОДВОДНАЯ ЛОДКА*
ПОДВОДНЫЕ РОБОТОТЕХНИКИ	*ЦИФРОВАЯ МОДЕЛЬ РЕЛЬЕФА*
ПОДВОДНЫХ РОБОТОТЕХНИКОВ	*ИССЛЕДОВАНИЯ МОРСКОГО ДНА*
ПОДВОДНЫХ МИССИЙ	*МИНИ-ПОДЛОДКА*
ПОДВОДНЫЕ МИССИИ	*МИНИ-ПОДЛОДКИ*
ПОДВОДНАЯ МИССИЯ	*ИПМТ ДВО РАН*
UNMANNED VEHICLE SYSTEM	*ДНО ОКЕАНА*
ДВФУ	*ДНЕ ОКЕАНА*
АКУСТИЧЕСКАЯ АНТЕННА	*ПОДВОДНЫЙ РЕЛЬЕФ*
АКУСТИЧЕСКИМИ АНТЕННАМИ	*САНПА*
ВОЕННЫЕ ВОДОЛАЗЫ	*LDUUV*
ВОЕННЫМ ВОДОЛАЗАМ	*ОБЗОР-600*
OPENROV	*ТЕТИС-ПРО*
ГЛУБОКОВОДНЫЕ	*ГИДРОЛОКАЦИОННЫЕ ИЗОБРАЖЕНИЯ*
ПОДВОДНЫЕ АППАРАТЫ	*ГИДРОЛОКАЦИОННЫХ ИЗОБРАЖЕНИЙ*
АВТОНОМНЫЕ ПОДВОДНЫЕ РОБОТЫ	*ГЛУБОКОВОДНЫЕ АППАРАТЫ*
ТНПА	*БЕСПИЛОТНЫЙ ПОДВОДНЫЙ АППАРАТ*
ПРОТИВОМИННАЯ БОРЬБА	*ПОДВОДНЫЙ АППАРАТ-РОБОТ*
ПРОТИВОМИННОЙ БОРЬБЫ	*НПА-РОБОТ*
ОБУЧЕННЫЕ ДЕЛЬФИНЫ	*МАГНЕТОМЕТРЫ*
КАРТА ДНА	*ПОДВОДНОЕ ВИДЕОНАБЛЮДЕНИЕ*
КАРТЫ ДНА	*ПОДВОДНОГО ВИДЕОНАБЛЮДЕНИЯ*
АНПА REMUS	*БЕСПИЛОТНИКИ*
СОНАРЫ	*ЦИФРОВАЯ МОДЕЛЬ РЕЛЬЕФА ДНА*
SEAFOX	*ЦМРД*
ПОДРЫВ ДРЕЙФУЮЩИХ МИН	*ПОДВОДНЫЙ ВУЛКАН*
ПОДРЫВА ДРЕЙФУЮЩИХ МИН	*KNIFEFISH*
АНПА МТ-88	*RECCE*
	АНПА AUV-150

Статистический анализ дайджеста показал следующую статистику для КТ.

АНПА => вес: 40	подводный аппарат => вес: 44
необитаемый подводный аппарат => вес: 39	подводный аппарат тнпа => вес: 29
дна анпа remus => вес: 38	телеуправляемый подводный аппарат => вес: 29
аппарат анпа remus => вес: 37	анпа и тнпа => вес: 27
аппарат анпа => вес: 37	морского дна => вес: 26
подводный аппарат анпа => вес: 36	подводный аппарат ипмт => вес: 25
аппарат анпа recce => вес: 36	unmanned underwater vehicle => вес: 25
автономный необитаемый подводный => вес: 35	тнпа или анпа => вес: 25
необитаемый подводный => вес: 35	подводный аппарат который => вес: 25
анпа remus => вес: 35	подводный аппарат для работы => вес: 24
дна анпа => вес: 34	ипмт дво ран => вес: 22
подводный аппарат => вес: 34	аппарат тнпа => вес: 22
морского дна => вес: 33	телеуправляемый подводный => вес: 22
морского дна анпа => вес: 32	аппаратов анпа => вес: 22
анпа обзор600 => вес: 31	аппаратов анпа gavia => вес: 22
анпа recce => вес: 30	unmanned underwater => вес: 22
подводных анпа => вес: 30	underwater vehicle => вес: 22
рельефа анпа => вес: 29	дво ран => вес: 21
автономный необитаемый => вес: 29	рельефа дна непосредственно => вес: 21
анпа и способен => вес: 29	
подводный аппарат обзор600 => вес: 27	
анпа => вес: 25	
анпа remus 100 позволяет => вес:	

25 разведки морского дна => вес: 23 рельефа морского дна => вес: 23	вмс говорится что анпа => вес: 21 gavia анпа => вес: 20 анпа gavia => вес: 20 displacement unmanned underwater => вес: 20 эхолот который => вес: 19 underwater vehicle lduuv => вес: 19

После получения списка КТ(полученный при анализе дайджеста) было проведено тематическое моделирование ПО с помощью метода LDA, который обрабатывал данные ключевые термины как отдельные слова.

Список тем полученных после тематического анализа.

В результате дальнейшей работы системных модулей (анализа дайджеста) были были выделены «производители АНПА», «покупатели АНПА», «виды АНПА», « модели АНПА», «оборудование», используемое для оснащения АНПА, объекты исследования АНПА

Описанная автоматическая технология семантического серфинга впервые предоставил возможность выбора и накопления больших объемов релевантных и качественных текстовых данных из Интернет по неограниченному количеству разнообразных предметных областей, заданных набором ключевых терминов.

Заключение

Данный алгоритм создания корпуса из ключевых фраз имеет ряд существенных преимущесттв в плане посткроулинговой обработки. Нам не

требуется тратить время и вычислительные ресурсы дополнительно на удаление идеальных дубликатов, шаблонов удалять HTML- коды, осуществлять проверку, находится ли ключевые фразы в языке перевода, осуществлять фильтрацию подозрительных страниц, таких как те, что генерируются автоматически из БД. Хотя заголовки могут все же попасть в корпус, но они в соответствии с принципами формирования корпуса, тоже могут являться ключевыми фразами. Также не требуется осуществлять работу по исключению документов с большой долей неязыковых материалов. Кроме того для корпуса, который предназначен для решения вышеозначенных задач не требуется никакого аннотирования. Все это экономит время , вычислительные и денежные ресурсы.

Единственно, что неплохо было бы добавить лемматизацию, те приведения словоформ к их нормальной (словарной) форме, а после этого исключение терминов-дубликатов, а также исключение служебных слов при формировании списка значимых терминов для дальнейшей работы по составлению тезаруса и АИППО.

Стоит отметить, что к настоящему времени обработана предметная область «Автономные необитаемые подводные аппараты» (АНПА) на русском и английском языках. В рамках данной предметной области обнаружено и обработано около 1 миллиона документов общим объемом свыше 120 ГБ, из которых выделены относящиеся к теме АНПА тексты

Развитие предложенной методики в будущем предполагается осуществить в следующих направлениях:

1) поиск текстов по заданной предметной области при помощи разрабатываемого аппарата АИППО, в результате чего будет сделан шаг к автоматическому построению онтологий предметной области;

2) смысловой поиск и построение аналитического отчета по заданной предметной области при помощи разрабатываемого аппарата АИППО;
3) распространение методики на многие иностранные языки;
4) использование аппарата АИППО для мониторинга предметной области.

Были проведены первые эксперименты по применению метода LDA к набору ключевых фраз из дайджеста. Были определены оптимальные параметры количества выявленных тем и числа терминов в теме. На основе дайджеста и результатов работы алгоритма LDA построен аналитический отчет по предметной области АНПА, содержащий разделы: «Общая характеристика АНПА», «АНПА зарубежного производства», «АНПА отечественного производства», «Объекты исследования АНПА», «Оборудование АНПА».

Список использованной литературы

1. [21] McEnery, T., and Wilson, A. (2001) Corpus Linguistics, 2nd edition. Edinburgh: Edinburgh University Press.
2. [10] Clark, A., Cussens, J., Sakas, W., Xantho, A. (eds.)(2005) Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition of ACL 05.New Brunswick: ACL.
3. [24] Sinclair, J. (2003) ‘Corpora for lexicography’, in A practical guide to lexicography, P. Van Sterkenberg (ed.) Amsterdam: John Benjamins.
4. [20] Manning, Ch., and Sch ¨ utze, H. (1999) Foundations of Statistical Natural Language Processing. Cambridge MA:MIT Press.
5. [3] Banko, M., and Brill, E. (2001) ‘Scaling to very very large corpora for natural language disambiguation’, Proceedings of ACL-01.
6. [19] Mair, Ch. (2003) ‘Tracking ongoing grammaticalchange and recent diversification in present-day standardEnglish: the complementary role of small and large corpora’, paper presented at the Annual ICAME Conference

7. [27] Turney, P. (2001) ‘Mining the Web for synonyms: PMIIR versus LSA on TOEFL’. Proceedings of ECML 2001,491-502.
8. [23] Sharoff, S. (2006) ‘Creating general-purpose corpora using automated search engine queries’, in [6].
9. [28] Ueyama, M. (2006) ‘Creation of general-purpose Japanese Web corpora with different search engine query strategies’, in [6].
10. [25] Storrer, A., and Beißwenger, M. (to appear) ‘Corpora of computer-mediated communication’, in Corpus linguistics: An international handbook, A. L ¨ udeling and M. Kyt ¨ o(eds.). Berlin: Mouton de Gruyter.
11. [15] Ghani, R., Jones, R., Mladenic, D. (2001) ‘Using the Web to create minority language corpora’. Proceedings of the 10th International Conference on Information and Knowledge Management
12. [5] Baroni, M., and Bernardini, S. (2004) ‘BootCaT: Boot strapping corpora and terms from the web’, Proceedings of the Fourth Language Resources and Evaluation Conference.
13. [14] Fletcher, B. (2004) ‘Making the Web more useful as a source for linguistic corpora’. In Corpus Linguistics in North America 2002, U. Connor and T. Upton (eds.) Amsterdam: Rodopi.
14. [11] Clarke, C., Cormack, G., Laszlo, M., Lynam, T., and Terra, E. (2002) ‘The impact of corpus size on questionanswering performance.’ Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval
15. [26] Thelwall, M. (2005) Creating and using Web corpora.International Journal of Corpus Linguistics 10(4), 517-541.
16. [13] Emerson, T., O’Neil, J. (2006) ‘Experience building a large corpus for Chinese lexicon construction.’ In [6].
17. [2] Meng W., Yu C., Liu K. L. Building Efficient and Effective Metasearch Engines // ACM Computing Surveys (CSUR). 2002. Vol. 34. No. 1. P. 48–89.

18. [7] Berjon R., Faulkner S., Leithead T., Navara E. D., O'Connor E., Pfeiffer S., Hickson I.HTML5: A Vocabulary and Associated APIs for HTML and XHTML // W3C Candidate Recommendation. 2013.
19. [8] Кузнецов Р. Ф. Извлечение значимой информации изweb-страниц с использованием предложений// RCDL'2006: Сб. тез. постерных докл. VIII Всерос. конф. СПб.: НУ ЦСИ, 2006. 274 с.
20. [9] Baumgartner R.Datalog-Related Aspects in Lixto Visual Developer // Datalog Reloaded. Lecture Notes in Computer Science. 2011. Vol. 6702. P. 145–160.
21. [10] Агеев М. С., Вершинников И. В., Добров Б. В. Извлечение значимой информации из web-страниц для задач информационного поиска// Интернет-математика2005. Автоматиче-ская обработка веб-данных. М., 2005. С. 283–301.
22. [11] Marathe M., Patil S. H., Garje G. V., Bewoor M. S. Extracting Content Blocks from Web Pages // International Journal of Recent Trends in Engineering, 2009. Vol. 2. No. 4. P. 62–64.
23. Charnine M.M., Kuznetsov I.P., Kozerenko E.B. Semantic Navigator for Internet Search.// Proceeding of International Conference on Machine Learning, 27-30, 2005 Las Vegas, USA, CSREA Press, pp 60-65, 2005.
24. Michael Charnine, Vladimir Charnine. Keywen Category Structure.// Wordclay, USA, 2008, pp.1-60 (монография).
25. Michael Charnine, "Keywen Automated Writing Tools", Booktango, USA, 2013, ISBN 978-1-46892-205-9.
26. Шарнин М.М., Кузнецов И.П. Автоматическое формирование электронных энциклопедий и справочных пособий по информации из сети "Интернет". //Системы и средства информатики. Вып.14, ИПИ РАН, 2004 г., с. 210-223.
27. M. M. Charnine, I. P. Kuznetsov, E. B. Kozerenko,. Technological peculiarity of knowledge extraction for logical-analytical systems.

WORLDCOMP’12 July 16-19, 2012. Las Vegas, USA.// CSREA Press, pp. 49-55, 2012.

28. Шарнин М.М., Кузнецов И.П. Особенности семантического поиска информационных объектов на основе технологии баз знаний. // «Информатика и ее применения» т.6, Вып. 2. 2012, стр. 47-56.2012.
29. M. M. Charnine, et al. Intelligent Tools for the Semantic Internet Navigator Design. // Труды конф. RCDL. Переяславль-Залесский. 2012. С. 274-283.
30. Шарнин М.М., Мацкевич А.Г., Кузнецов И.П. Технология извлечения структур знаний с использованием аппарата расширенных семантических сетей. // Журнал «Искусственный интеллект», НАН Украины, 2012. Том 4.
31. М. М. Шарнин, Н. В. Сомин, И. П. Кузнецов, Ю. И. Морозова, И. В. Галина, Е. Б. Козеренко. Статистические механизмы формирования ассоциативных портретов предметных областей на основе естественно-языковых текстов больших объемов для систем извлечения знаний. // Информатика и её применения. 2013. Т.7. №2. Стр.92–99.
32. M. Charnine, A. Petrov, I. Kuznezov. Association-Based Identification of Internet User Interests. // Proceedings of the 2013 International Conference on Artificial Intelligence (ICAI 2013). V. I. WORLDCOMP’13. July 22-25, 2013. Las Vegas Nevada. USA. CSREA Press. С. 77-81.
33. M. Charnine, V. Protasov. Optimal Automated Method for Collaborative Development of Universiry Curricula. // Proceedings of the 2013 International Conference on Artificial Intelligence (ICAI 2013). V. I. WORLDCOMP’13. July 22-25, 2013. Las Vegas Nevada. USA. CSREA Press. С. 96-100.
34. Elisabeth Bacon, George Hagel, Michael Charnine, Richard Foggie, Brian Kirk, Igor Schagaev, and George Kravtsov. WEDUCA: Web-enhanced

design of university curricula. // Proceedings of the FECS'13: The International Conference on Frontiers in Education: Computer Science and Computer Engineering. July 22-25, 2013. Las Vegas, Nevada, USA. CSREA Press. С. 288-294.

35. Шарнин М.М., Протасов Владислав Иванович Оптимальный автоматизированный метод коллективной разработки учебных программ ВУЗов //Труды XVIII международной конференции «Технологии будущего для человечества» (СРТ’2013), Ларнака, Кипр, 12-19 мая 2013 //с.312-314//Издательство Института физико-технической информатики (ИФТИ)
36. Шарнин М.М., Рыков Владимир Васильевич, Клименко Станислав Владимирович Автономные необитаемые подводные аппараты: автоматическое формирование ассоциативного портрета предметной области//Труды Международной научной конференции MEDIAS.//Изд-во Института физико-технической информатики (ИФТИ)
37. Шарнин М. М., Петров А.В., Кузнецов И.П. “Методика учёта интересов пользователя при работе в сети Internet на основе его профиля и ассоциативных связей”// Труды XV Всероссийской научной конференции RCDL’2013 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Ярославль, 14-17 октября 2013. Ярославский государственный университет им. П.Г. Демидова (ЯрГУ), 2013. С. 86-90.
38. Bacon E., Hagel G., Charnine M., Foggie R., Kirk B., Schagaev I., Kravtsov G. WEDUCA: Web-enhanced design of university curricula // Proceedings of the FECS'13: The International Conference on Frontiers in Education: Computer Science and Computer Engineering, July 22-25, 2013, Las Vegas, Nevada, USA, CSREA Press, 2013. P. 288-294.

39. O. Zolotarev, M. Charnine, A. Matskevich. "Conceptual Business Process Structuring by Extracting Knowledge from Natural Language Texts"//Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014), vol.I, WORLDCOMP'14, July 21-24, 2014. Las Vegas Nevada, USA. CSREA Press, pp.82-87.
40. M. Charnine, N. Somin, V. Nikolaev. "Conceptual Text Generation Based on Key Phrases"//Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014), vol.I, WORLDCOMP'14, July 21-24, 2014. Las Vegas Nevada, USA. CSREA Press, pp.639-643.
41. V. Protasov, M. Charnine, E. Melnikov. "The Crowdsourcing Linguistic Technology for Experts Assessment"//Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014), vol.I, WORLDCOMP'14, July 21-24, 2014. Las Vegas Nevada, USA. CSREA Press, pp.656-661.
42. Морозова Ю. И., Козеренко Е. Б., Шарнин М. М. «Методика извлечения пословных переводных соответствий из параллельных текстов с применением моделей дистрибутивной семантики» // Журнал: Системы и средства информатики, 2014. Т. 24. Вып. 2. С. 131–142.
43. Морозова Ю. И., Козеренко Е. Б., Будзко В. И., Кузнецов К. И., Шарнин М. М. Семантическая структуризация текстовых знаний для систем аналитического мониторинга больших объемов информации в социальной сфере // Журнал : «Системы высокой доступности», 2014. Т. 10. № 3. С. 21–35.
44. Борисов Т. Н., Клименко С. В., Рыков В. В., Хламов М. А., Шарнин М. М. Построение онтологий для обеспечения информационной поддержки лиц, принимающих решения по проблеме распространения отравляющих веществ из контейнеров и корпусов боеприпасов подводного захоронения // Ситуационные центры и информационно-

аналитические системы класса 4i для задач мониторинга и безопасности (SC-IAS4i-VRTerro2013): Труды II Международной научной конференции (Протвино, 25–29 ноября 2013). – Протвино: ИФТИ, 2013. С. 255–261. (В библиографии 2013 года не отражена.)

45. Протасов В. И., Шарнин М. М., Мельников И. Е. “Применение технологии самоуправляемого краудсорсинга для сертификации экспертов” // Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности (SC-IAS4i-VRTerro2013): Труды II Международной научной конференции (Протвино, 25–29 ноября 2013). – Протвино: ИФТИ, 2013. С. 262–266.
46. О.В.Золотарев, М.М.Шарнин "Методы извлечения знаний из текстов естественного языка и построение моделей бизнес-процессов на основе выделения процессов, объектов, их связей и характеристик"//Труды IXX международной конференции CPT’2014, Ларнака, Кипр, 11-18 мая 2014, Издательство Института физико-технической информатики (ИФТИ)
47. Шарнин М. М., Родина И. В. Механизмы формирования лингво-статистического портрета предметной области мониторинга общественного мнения для технологии обработки «Больших данных» // Социальный компьютинг: основы, технологии развития, социально-гуманитарные эффекты (ISC-14): Материалы III Международной научно-практической конференции (Москва, 18–19 сентября 2014): Сборник статей и тезисов. – М.: МГГУ им. М. А. Шолохова, 2014. С. 66–70. [Электронное издание] http://mggu-sh.ru/sites/default/files/sb_2014.pdf.
48. M.Charnine, S.Klimenko, "Measuring of “Idea-based” Influence of Scientific Papers"Proceedings of the 2015 International Conference on Information Science and Security (ICISS 2015), December 14-16, 2015, Seoul, South Korea, pp.160-164.

49. М.М. Шарнин, О.В. Золотарев, Н.В. Сомин "Извлечение и обработка знаний из неструктурированных текстов деловой сферы и социальных сетей"//4-я международная научно-практическая конференция "Социальный компьютинг: основы, технологии развития, социально-гуманитарные эффекты", Москва, МПГУ, 22-24 октября.
50. Шарнин М.М., Шагаев И., Протасов В.И., Родина И.В., Золотарев О.В., Попова О.А., "Использование веб-семантики для совершенствования образовательных программ вузов",Вестник МГГУ им. М.А.Шолохова, Филологические науки, 2015, № 2, с.97-112 // http://www.academia.edu/14982851/ // http://mggu-sh.ru/sites/default/files/sharninshagaevprotasovrodinazolotarevpopova_0.pdf // http://mggu-sh.ru/vestnik/vestnik-filologicheskie-nauki-2015-no-2-0
51. Galina, M. Charnine, N. Somin, V. Nikolaev, Yu. Morozova, O. Zolotarev, “Metod for generating subject area associative portraits: different examples”, Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015), WORLDCOMP’15, July 27-30, 2015, Las Vegas Nevada, USA, v.I, pp.288-294,ISBN: 1-60132-405-7, 1-60132-406-5 (1-60132-407-3).
52. O. Zolotarev, M. Charnine, A. Matskevich, K. Kuznetsov,“Business Intelligence Processing on the Base of Unstructured Information Analysis from Different Sources Including Mass Media and Internet”, Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015), WORLDCOMP’15, July 27-30, 2015, Las Vegas Nevada, USA, v.I, pp.295-299,ISBN: 1-60132-405-7, 1-60132-406-5 (1-60132-407-3).
53. M. Charnine, N. Somin, S. Klimenko, V. Ezhela, "Linguistic Approach to Scientometrics", Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015), WORLDCOMP’15, July 27-30, 2015, Las Vegas, Nevada, USA, v.II, pp.812-817,ISBN: 1-60132-405-7, 1-60132-406-5 (1-60132-407-3). http://www.worldcomp-

proceedings.com/proc/proc2015/ICAI15_Final_Edition/ICAI_Contents_Vol_2.pdf

54. В.В. Ежела, С.В. Клименко, А.Н. Райков, М.М. Шарнин,"Семантический подход к оценке качества научных публикаций",Научно-техническая информация (НТИ), Сер.2, Информационные процессы и системы, 2015, № 7, С.13-18, ISSN 0548-0027.
55. V.V. Ezhela, S.V. Klimenko, A.N. Raikov, M.M. Sharnin,"A semantic approach to the evaluation of the quality of academic publications", Automatic Documentation and Mathematical Linguistics 07/2015, 49(4), pp.117-121, DOI: 10.3103/S0005105515040020.
56. Шарнин М.М., Сомин Н.В., Галина И.В., Родина И.В., Николаев В.Г.Осмысленные генеранты: метод порождения текстов повышенной информативности из релевантных фраз. // Труды XIX международной конференции СРТ2014. Ларнака, Кипр. 12-18 мая 2014. //Изд-во Института физико-технической информатики (ИФТИ), 2015. С. 87-91.http://elibrary.ru/item.asp?id=24262655(РИНЦ)
57. Борисов Т.Н., Бронецкий А.Е., Клименко С.В., Рыков В.В., Шарнин М.М. Автономные необитаемые подводные аппараты: автоматическое формирование ассоциативного портрета предметной области // Сборник трудов Международной конференции «Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности» SCVRT2013-14, Протвино, Парк Дракино, 25--28 ноября 2013-2014 гг. Изд. ИФТИ, Москва-Протвино, 2013-2014, ISBN 978-5-88835-027-0, С.38-43
58. Anton Zimmerling, "Parametrizing verb second languages and clitic second languages",Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015), WORLDCOMP'15, July 27-30, 2015, Las Vegas

Nevada, USA, v.I, pp.281-287, ISBN: 1-60132-405-7, 1-60132-406-5 (1-60132-407-3).