# Семантический анализ Big Data в задаче прогнозирования инновационного развития энергетической инфраструктуры РФ

А.Н. Копайгородский, Е.П. Хайруллина, И.И. Хайруллин
Институт систем энергетики им. Л.А. Мелентьева СО РАН
kopaygorodsky@isem.irk.ru | Lena-Skoklenyova@yandex.ru | IIlyaKhairullin@gmail.com

*В статье рассматривается применение методов семантического анализа и обработки естественного языка для поддержки исследований и прогнозирования инновационного развития энергетической инфраструктуры РФ. Рассмотрены существующие методы и подходы к организации мониторинга технологических решений и инновационных научных разработок. Для автоматизации мониторинга авторами предлагается применение методов обработки естественного языка (NLP). Семантический анализ и интеграция знаний основаны на системе онтологий. В работе представлены основные методы и подходы к построению инфраструктуры для обработки открытых Больших данных (Big Data). Применение предложенных методов позволяет повысить качество научных исследований в этой области и сделать их более эффективными.*

***Ключевые слова:*** *научно-технологическое прогнозирование, семантический анализ, обработка естественного языка, Большие данные, поддержка научных исследований.*

# Semantic Analysis of Big Data in the Problem of Forecasting the Innovative Development of the Energy Infrastructure of the Russian Federation

A.N. Kopaygorodsky, Khayrullina E.P., Khayrullin I.I.
Melentiev Energy Systems Institute SB RAS

*The article discusses the use of methods of semantic analysis and natural language processing to support research and forecasting the innovative development of the energy infrastructure of the Russian Federation. The existing methods and approaches to the organization of monitoring of technological solutions and innovative scientific developments are considered. To automate monitoring, the authors propose the use of natural language processing (NLP) methods. Semantic analysis and knowledge integration are based on a system of ontologies. The paper presents the main methods and approaches to building an infrastructure for processing open Big Data. Application of the proposed methods makes it possible to improve the quality of scientific research in this area and make them better.*

*Key* ***words****: scientific and technological forecasting, semantic analysis, natural language processing, Big data, scientific research support.*

## 1. Введение

В последние годы активно развиваются методы обработки естественного языка и семантического анализа, что позволяет использовать их для решения задач научно-технологического прогнозирования, в том числе и в области энергетических технологий. Применение методов Data Science в высокотехнологичных отраслях позволяет значительно повысить эффективность принимаемых управленческих решений. Подобный подход к управлению в отдельных организациях различного уровня принято называть «управление, основанное на данных». При реализации такого подхода в организациях создаются специализированные информационно-аналитические отделы под руководством CDO (Chief Digital Officer / Chief Data Officer), мнение которого является ключевым в вопросах развития компании, определения новых бизнес-возможностей, обеспечении выхода на новые сегменты рынка, вывода принципиально новых продуктов, сервисов и услуг. Основным источником для принятия управленческих решений являются результаты анализа информации и данных, собранных из различных источников, для выявления существующих и зарождающихся трендов и сокращения времени реакции на них, что уменьшает материальные потери, увеличивает прибыль и обеспечивает устойчивое развитие компании. Игнорирование технико-экономических трендов может оказаться фатальным или нанеси существенный финансовый ущерб компании. Применение такого подхода актуально при решении задач прогнозирования и организации мониторинга инновационных технологических решений в любой отрасли народного хозяйства, а особенно для развития энергетической инфраструктуры России, поскольку последняя является базовой для функционирования других отраслей, результаты деятельности которых, в конечном итоге, направлены на улучшение качества жизни населения. Прогнозные методы развития энергетической отрасли, основанные на традиционных математических моделях и программных комплексах, не всегда эффективны в условиях неопределенности и отсутствия необходимой достоверной информации для имеющихся моделей. Актуальным является создания инструментария, облегчающего работу экспертов, и инструментария, выполняющего предварительную обработку информации, анализируемой экспертами, с привлечением интеллектуальных методов семантического анализа, машинного обучения и технологии Big Data.

## 2. Анализ Инновационных технологических решений

На протяжении последних десятилетий международная рабочая группа ведущих ученых из США, Европы и Восточной Азии координирует исследования по анализу перспективных технологий (Future Oriented Technology Analysis) [1]. Исследования выполняются в таких областях как технологический мониторинг (technology monitoring, watch, alerts), системное изучение технологий и конкурентной среды (technical and competitive intelligence), научно-технологическое прогнозирование (technological forecasting), оценка стратегических направлений технологического развития и последствий технологических изменений (technology assessment), разработка технологических дорожных карт (technology roadmapping) и технологический форсайт (technology foresight).

Кроме того, исследования в области научно-технологического прогнозирования выполняются как в рамках отдельных научно-исследовательских программ, так и специализированными подразделениями ведущих мировых компаний. Например, в рамках исследовательского проекта «Технологическое прогнозирование с использованием интеллектуального анализа данных и семантики / Technological Forecasting using Data Mining and Semantics», выполняемым Массачусетским технологическим университетом, разрабатываются основанные на семантических технологиях методы с целью анализа массива научно-технической информации из электронных ресурсов, в том числе и в сфере энергетики [2].

До недавнего времени для изучения энергетики РФ в основном использовались следующие базовые методы технологического прогнозирования:

1. Экспертные оценки (метод Delphi и Foresight), которые позволяют получать простые и быстрые решения, однако этим решениями свойственны такие плохие качества как субъективность и низкая обоснованность.
2. Технологический анализ в отдельных энергетических компаниях, основанный на эконометрике, мониторинге, базах данных разработок, оперативном анализе проблем и существующих предложений.
3. Системный анализ технологии, который требует достаточно много времени и усилий высококвалифицированных научных команд, что делает этот подход очень дорогим и не всегда осуществимым.

В процессе построения научно-технического прогноза требуется учитывать влияние таких основных факторов и ограничений как:

1) необходимость поддержки устойчивого научно-технического развития;
2) ускорение технологических изменений;
3) усиление влияния прогресса в области науки и технологий на социально-экономические процессы;
4) рост сложности объектов, контрагентов и, как следствие, объединяющих их систем;
5) усиление междисциплинарных эффектов, влияния друг на друга смежных областей знания, в том числе конвергенция технологий;
6) необходимость систематического адаптивного прогноза и механизмов скользящего планирования при непрерывной актуализации научно-технической информации.

В процессе своей работы исследователи Института систем энергетики им. Л.А. Мелентьева Сибирского отделения Российской академии наук (ИСЭМ СО РАН) испытывают в некоторой степени трудности с актуальными данными и знаниями для выполнения высокоточного научно-технологического прогнозирования развития энергетики. Также следует отметить, что технологический прогноз развития энергетики РФ до 2035 года рассматривает три разных сценария развития мировой энергетики [3]. Вероятность каждого из предложенных сценариев не может быть правильно оценена, что приводит к высокой неопределенности в стратегическом планировании инновационного развития из-за высоких инвестиционных рисков в секторе энергетики. Показательным является множество различных событий, произошедших в 2020 году и значительно повлиявших на мировую энергетику: при составлении сценариев ее развития не возможно было учесть совокупность таких факторов и рисков, которые при своей реализации привели к беспрецедентному падению цен на нефть марки West Texas Intermediate (WTI) до -$38 за баррель в апреле с поставкой в мае 2020 года. Такое ценообразование и состояние мировой энергетики негативно сказалось на развитие инвестиционных и научно-исследовательских проектов. Улучшение оценки технологического развития и скользящее прогнозирование позволит сосредоточиться на усилении выявленных технологий с лучшим эффектом в будущем. Таким образом, предоставление исследователям актуальной информации и знаний является важной частью выработки экспертных оценок новых технологических решений в энергетике. В качестве источников информации могут использоваться текстовые данные из открытых государственных систем и других источников. Например, чтобы улучшить качество результатов и повысить вероятность прогнозов развития энергетики, исследователи могут обратиться к базам патентов и изобретений, научных публикаций или данным новостных лент и др., однако обработка такого большого массива данных является сложной и трудозатратой.

## 3. Обработка текстов на естественных языках

Существует достаточно много методов обработки естественного языка, используемых в информационных системах. Все методы можно условно разделить на две большие группы: статистические и лингвистические методы. [4]

Статистические методы опираются на анализ частотности встречаемости слов: подсчет количества вхожде-

ний слов в различные фрагменты, распределение частотности по документам и пр. Лингвистический анализ, напротив, основывается на выявлении отдельных слов, анализе их морфологических признаков, синтаксическом и семантическом анализе текстовых фрагментов. В результате выполнения синтаксического анализа входные текстовые данные преобразуется в иерархические структуры последовательностей лексем языка, что позволяет значительно упростить дальнейший анализ. При выполнении этапа семантического анализа выделяются семантические отношения между отдельными лексемами, формируется семантическое представление отдельных последовательностей лексем.

Важную роль для улучшения результатов анализа и уменьшения области поиска играет технология стемминга. Стемминг [5] позволяет выявить основу слова, связать множество форм одного и того же слова между собой, что позволяет значительно проще обрабатывать массивы текстовых данных. Важной особенностью стемминга является его зависимость от языка, поскольку правила словообразования для каждого естественного языка, обычно, являются специфичными. Наиболее лучшие результаты показывают стемминг на основе предварительно рассчитанных таблиц и словарей. Однако применение исключительно такого подхода имеет некоторые сложности: невозможность определения основы нового слова, которое не было предварительно обработано. Наиболее рациональным при построении стеммеров является комбинация алгоритмического и словарного подхода, что позволяет с одной стороны получить достаточно хорошие результаты, а также сохраняется возможность вероятностного анализа неизвестных стеммеру слов.

Одной из наиболее распространенных статистических мер анализа текстов является TF-IDF [6], где TF – это Term Frequency, а IDF – это Inverse Document Frequency. Мера TF-IDF позволяет оценить важность слова в контексте документа, который входит в коллекцию документов (корпус). Значение меры TF-IDF для слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах собранного корпуса. Мера TF-IDF позволяет достаточно легко отделить слова общей лексики от специфичных терминов: для слов общей лексики мера будет стремиться к нулю, тогда как для редких терминов быстро возрастать. Однако стоит отметить, что неправильно подобранный корпус в значительной степени влияет на меру TF-IDF.

В последнее время все большее применение находит статистический метод анализа word2vec [7, 8]. Этот алгоритм позволяет представить слова в виде многокомпонентного вектора. Word2vec был разработан компанией Google в 2013 году и нашел отражение в коммерческих проектах многих компаний. Word2vec базируется на совокупности моделей на основе искусственных нейронных сетей, предназначенных для получения векторных представлений слов на естественном языке. Word2vec использует большую коллекцию текстовых документов в качестве входных данных и сопоставляет каждому слову N-вектор. При обучении модели word2vec исследователи задаются размерностью векторов, типичными применяемыми размерностями являются от 100, но обычно используется 300. Полученные вектора позволяют эффективно вычислять в векторном пространстве семантическую близость как отдельных слов, так и словосочетаний.

Наиболее распространены в области анализа данных два языка: Python и R. Язык R разрабатывался как инструмент статистической обработки и анализа данных с открытым кодом. Язык R, как и Python, хотя и является отчасти примитивным, но при этом имеет достаточно большое множество разнообразных библиотек, значительно упрощающих решение многих типовых задач в области анализа и обработки данных. Популярность языка Python обеспечивается с одной стороны поддержкой развитого математического аппарата, а с другой дополнительными библиотеками, значительно расширяющими возможности языка. Одной из наиболее прогрессивных библиотек для анализа текстов на Python является библиотека Natural Language Toolkit (NLTK) [9]. Библиотека NLTK включает множество инструментов для управления корпусами и их анализа:

- Токенизация (разбиение) по предложениям и словам. Данные методы позволяют разделять входной текстовый поток как на предложения, так и выделить отдельные слова. Такое разделение является исключительно важным и необходимо для любых дальнейших работ по анализу текста.
- Лемматизация и стемминг. Эти методы обеспечивают приведение всех слов входного потока к нормальной словарной форме, что позволяет уменьшить область анализа, увеличивая качество и скорость обработки текста.
- Методы векторизации и частотного анализа: мешок слов (Bag of Words – BoW) и TF-IDF. Мешок слов позволяет строить векторы для текстовых данных по заранее заданному словарю, с указанием в качестве значения соответствующих компонент встречаемость элементов словаря в тексте. Основным недостатком наряду с простотой реализации алгоритма является сложность в определении элементов заданного словаря, поскольку увеличение их количества приводит к разреженности в результирующих векторах, но при этом увеличивает их информативность.

При решении задачи семантического анализа для прогнозирования инновационного развития энергетической инфраструктуры применялись как статистические, так и лингвистические методы обработки текстов. После построения терминологического словаря предметной области на основе билингвистической онтологии для его отдельных элементов рассчитываются векторные представления. При обработке массивов тек-

стовых данных на первом этапе выполняется фильтрация и отделение слов общей лексики и других подобных языковых элементов, а на втором выполняется классификация оставшихся элементов и семантическое сопоставление с концептами онтологий. Далее полученные характеристики каждого документа обрабатываются статистически с целью выявления общих тенденций и закономерностей в данных.

## 4. Семантический анализ Big Data

При применении методов Large Data Analytics (LDA) к глобальным источникам данных о науке, технологиях и инновациях можно определить существующие и выявить новые тенденции развития, а также предвидеть технологические прорывы путем всестороннего понимания непрерывных инновационных процессов [10]. Новый подход к извлечению, накоплению и анализу знаний и данных позволяет улучшить качество получаемых научных результатов в задаче прогнозирования инновационного развития энергетической инфраструктуры [11].

В качестве источников информации для составления прогнозов используются открытые данные (Linked Open Data) из государственных информационных систем, а также из некоторых коммерческих систем, содержащие потенциально интересную информацию. Примерами таких систем являются базы научных публикаций, проводимых НИР, результатов интеллектуальной деятельности и др., которые, как правило, придерживаются определенной структуры публикуемых данных, и поэтому могут быть обработаны с применением программных адаптеров. Кроме того, для поиска неструктурированной, но потенциально интересной для исследователей информации, могут использоваться Интернет-поисковые системы с предварительным анализом, классификацией и оценкой найденной информации. Исходя из этого, все источники данных можно разделить на две категории по подходам к извлечению и обработке информации: структурированные и неструктурированные. Структурированные источники могут предоставлять информацию в соответствии с определенными структурами и, как правило, имеют API для организации программного доступа. Неструктурированные источники в первую очередь ориентированы на использование людьми, и поэтому, как правило, проиндексированы популярными Интернет-поисковыми системами. Сканирование источников информации на регулярной основе позволяет не только наполнять хранилище знаний, но и отслеживать динамику изменения качественных и количественных показателей на основе когнитивных моделей и Байесовских сетей доверия.

Семантическая интеграция данных и знаний основывается на использовании общего понятийного базиса и привязки (сопоставления) данных на его основе. Для задания такого базиса используется онтологическое пространство, включающее совокупность онтологий. Разработанная система онтологий включает онтологии ТЭК, отраслей энергетики и отдельных энергетических технологий, исследований энергетики, и является основным компонентом интеллектуальной системы, на основе которого выполняется настройка предметной области и интеграция знаний.

Поскольку анализ поступающей информации требует значительное время ввиду большого объема, было принято решение выделении двух процессов: процесса информационного наполнения хранилища и процесса использования наполненного хранилища исследователями для решения прикладных задач. [12]

Подсистема информационного наполнения состоит из модулей поиска (ферм краулеров) и модулей анализа и классификации. При этом один модуль анализа и классификации может обслуживать несколько модулей поиска, а несколько однотипных модулей поиска могут работать параллельно с информационными ресурсами одного класса. Модуль анализа при проведении классификации текстовых фрагментов опирается на данные извлеченные из билингвистической онтологии, задающей терминологическую специфику прогнозирования инновационного развития энергетики. Многие источники открытых данных, агрегирующие информацию о научных статьях и разработках, поддерживают экспорт записей в формате Research Information Systems (RIS). Формат RIS предназначен для обмена метаданными между различными научно-исследовательскими системами и содержит описание отдельных ресурсов (как правило, научных публикаций) в разрезе до 79 параметров, основными из которых являются название, информация об авторах, тип и дата публикации. Для анализа информации, размещенной на различных сайтах в сети Интернет, на первом этапе необходимо получение ссылок на такие сайты. Наиболее эффективным решением с точки зрения авторов было воспользоваться базами широко распространенных и хорошо известных информационно-поисковых систем, выполнивших индексацию информационных ресурсов в сети Интернет. Наиболее распространенными в российском сегменте сети являются Яндекс, Google и Bing. Мировым лидером поиска информации является Google. Работа краулеров основана на извлечении первичной информации из баз данных поисковых систем Яндекса, Google и Bing: на этом этапе можно получить адрес ресурса, заголовок и небольшое описание, а затем выполнить детальный анализ содержания ресурса, полученного по найденному адресу. Были разработаны специализированные инструментальные средства для поиска инициирующих ссылок по ключевым словам в базах данных поисковых систем Google, Bing, Яндекс с сохранением результатов в RIS-формате. Поиск проводился в разрезе 740 ключевых слов и выражений на русском и английском языках. Классификация текстовых фрагментов, извлеченных из документов краулерами, выполнялась на основе морфологического анализа, который

выполнялся реализованным на Python модулем с применением библиотеки NLTK.

В процессе сбора данных и подготовки к исследованиям и было обработано более 2,5 миллионов первичных документов полученных из различных открытых источников, выявлено и классифицировано более 500 тысяч документов на основе построенной системы онтологий, в том числе более 150 тысяч научных статей по системам энергетики (понятие “Power/Energy System”), проиндексированными различными международными системами цитирования и опубликованными в 2009-2020 годах, и о 220 тысячах действующих патентах, выданных в разных странах, применение которых возможно в области энергетики.

Одной из ключевых проблем после сбора информации является обеспечение эффективного хранения и доступа к ней. В результате поиска информации было идентифицировано 500 тысяч элементов и получено около 10 Гб текстовых данных. Данные в текстовом виде хранятся наиболее компактно и не содержат «неиспользуемых артефактов», кроме того к ним могут быть применены алгоритмы сжатия (архивирования), что дополнительно снизит их объем, но в любом случае невозможно организовать достаточно эффективный поиск по текстовым данным без использования дополнительных структур. Вследствие этого были выделены метаданные, для хранения и обработки которых была использована СУБД, а размер базы данных составил 1,7 Гб. Для ускорения доступа к метаданным собранных документов были построены специальные индексы и вынесены в отдельную базу данных, размер которой составил уже 270 Мб для 500 тысяч элементов. Таким образом, была построена иерархическая структура для хранения и обработки информации: «индексы – метаданные – текстовые данные – внешние ресурсы». Относительно компактные индексы позволяют быстро выполнить поисковые запросы, идентифицировать запрашиваемые пользователями элементы, затем выполнить доступ к конкретным метаданным отобранных элементов, а в случае необходимости обеспечить доступ как к сохраненной внутри системы текстовой информации, так и к опубликованной информации в сети Интернет. Обработка информации исключительно по индексам позволяет получать некоторые статистические оценки и строить временные ряды без необходимости обработки всего информационного массива.

## 5. Заключение

Семантический анализ Больших данных и методы обработки естественных языков и могут успешно применяться для прогнозирования инновационного развития энергетической инфраструктуры. Использование интеллектуальной информационной системы позволяет сократить время на разработку аналитических отчетов о перспективных исследованиях в области энергетики. Применение представленных методов и инструментальных средств, подходов и технологий позволит облегчить подготовку и повысить обоснованность опережающих рекомендаций и решений в области стратегического развития энергетики, а также обеспечить организацию мониторинга инновационных научно-технических решений и технологий в энергетике, включая их оценку их эффективности и реализуемости с учетом особенностей и потребностей экономики.


## Благодарность

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-07-00994.


## Литература


[1] C. Cagnin, M. Keenan, R. Johnston, F. Scapolo, R. Barre. Future-oriented technology analysis. strategic intelligence for an innovative economy. – Springer, 2008. – 170 p., DOI: 10.1007/978-3-540-68811-2.

[2] Woon, W.L., Zeineldin, H., Madnick, S. Bibliometric analysis of distributed generation // Technol. Forecast. Soc. Change, vol. 78(3), 2011. – pp. 408–420. – http://hdl.handle.net/1721.1/98911

[3] Прогноз научно-технологического развития отраслей топливно-энергетического комплекса России на период до 2035 года. Министерство энергетики Российской федерации, 2016. –http://minenergo.gov.ru/system/download-pdf/6365/66647

[4] В.В. Диковицкий, М.Г. Шишаев. Обработка текстов естественного языка в моделях поисковых систем / Труды Кольского научного центра РАН, 2010, №3. – С. 29-34.

[5] J.B. Lovins. Development of a Stemming Algorithm // Mechanical Translation and Computational Linguistics. – 1968. – T. 11.

[6] Jones K.S. A statistical interpretation of term specificity and its application in retrieval / Journal of Documentation: журнал. – MCB University: MCB University Press, 2004. – Vol. 60, no. 5. – P. 493-502. – ISSN 0022-0418.

[7] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR. – 2013a.

[8] Mikolov T., Yih W.-T., Zweig G. Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT. – 2013b.

[9] S. Bird, E. Klein, E. Loper. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. – http://www.nltk.org/book/

[10] T.U. Daim, D. Chiavetta, A.L. Porter, O. Saritas. Anticipating future innovation pathways through large data analysis // Innovation, technology, and knowledge management. – Springer, 2016, – 360 p., DOI 10.1007/978-3-319-39056-7.

[11] A. Kopaygorodsky. Methods and technologies of building an intelligent service for energy technology forecasting // International Scientific Journal Industry 4.0, 2018, Issue 5. – Scientific Technical Union of Mechanical Engineering "Industry 4.0". – Pp. 223-228. ISSN: 2534-8582

[12] A. Kopaygorodsky. Technology of Application of Software Tools for Energy Technology Forecasting // Proceedings of the 21st International Workshop on Computer Science and Information Technologies (CSIT 2019). Part of series Atlantis Highlights in Computer Sciences, vol. 3. 2019. – Atlantis Press. – Pp. 267-272. ISBN 978-94-6252-868-0, DOI: 10.2991/csit-19.2019.47.