УДК 004.82:004.912:001.18:620.9

ПРОЕКТИРОВАНИЕ И РЕАЛИЗАЦИЯ ИНСТРУМЕНТАЛЬНЫХ СРЕДСТВ ДЛЯ СЕМАНТИЧЕСКОГО АНАЛИЗА БОЛЬШИХ ДАННЫХ О НАУЧНЫХ И ТЕХНОЛОГИЧЕСКИХ РЕШЕНИЯХ В ОБЛАСТИ ЭНЕРГЕТИКИ

Копайгородский Алексей Николаевич
к.т.н., ведущий специалист отдела «Системы искусственного интеллекта в энергетике»
e-mail: kopaygorodsky@isem.irk.ru
Хайруллина Елена Павловна
аспирант отдела «Системы искусственного интеллекта в энергетике»
e-mail: lena-skoklenyova@yandex.ru
Институт систем энергетики им. Л.А. Мелентьева СО РАН
664033 г. Иркутск, ул. Лермонтова 130.

Аннотация. В статье рассмотрены подходы к проектированию и реализации отдельных компонентов инструментальных средств для семантического анализа извлекаемой из открытых источников информации о научных и технологических решениях в области энергетики. Рассмотрена структура билингвистической онтологии, позволяющая решать задачу классификации информации с учётом ее представления в различных языках и синонимии. Рассмотрен подход к поиску и обработке информации из открытых источников, основанный на применении разработанных авторами средств семантического анализа, реализация которых выполнялась на Python с использованием библиотеки Natural Language Toolkit.

Ключевые слова: научно-технологическое прогнозирование, семантический анализ, классификация текстовых документов, билингвистическая онтология.

Цитирование: Копайгородский А.Н., Хайруллина Е.П. Проектирование и реализация инструментальных средств для семантического анализа Больших данных о научных и технологических решениях в области энергетики // Информационные и математические технологии в науке и управлении. 2021. № 4 (24). С. 100-110. DOI: 10.38028/ESI.2021.010.

Введение. Научно-технологическое прогнозирование и организация мониторинга инновационных технологических решений играют важную роль в современном постиндустриальном обществе. Активное развитие информационно-телекоммуникационных технологий в значительной мере влияет на сокращение времени выхода новых инновационных разработок на рынок за счет ускорения передачи научных знаний и реализации на их основе новых производственных технологий с последующим выпуском продукции. Поиск новых технологических решений, предсказание перспективы их развития в жесткой конкурентной среде позволяют приобрести значительные преимущества не только отдельным компаниям, но и целым отраслям национальных экономик на глобальном рынке. Систематический характер таких поисковых работ в области энергетики требует развития методов семантического анализа Больших данных (Big Data) для выработки оценок и опережающих рекомендаций, а также создания новых инструментальных средств для поддержки этой деятельности. Необходимость развития методов анализа и обработки Big Data с помощью интеллектуальных информационных систем подчеркивается Национальной технологической инициативной (НТИ) России, а их применение в области энергетики соотносится с рынком EnergyNet, что находит отражение в «Дорожной карте», одобренной в 2016 году Президиумом Совета при Президенте Российской Федерации по модернизации экономики и инновационному развитию России.

Применение интеллектуальных методов семантического анализа, машинного обучения и технологии Big Data и создание инструментария, выполняющего

предварительную обработку массивов информации, позволяют значительно облегчить работу экспертных групп при решении поставленной задачи. Источниками информации могут выступать Открытые данные (Open Data) и Большие данные. Кроме того, эксперты могут эффективно использовать для «экспресс-анализа» собранной информации методы семантического моделирования [1-3], разработка которых ведется в Отделе систем искусственного интеллекта в энергетике Института систем энергетики им Л.А. Мелентьева Сибирского отделения РАН (ИСЭМ СО РАН).

Управление инновациями и поиск технологических решений. В последние десятилетия активное развитие получили интеллектуальные подходы и методы поддержки принятия решений, в том числе и в области планирования и управления инновационным развитием [4, 5]. С начала 2000-х годов была сформирована международная рабочая группа из ведущих ученых США, Европы и стран Восточной Азии, которая координирует исследования по перспективному анализу научно-технологического развития (Future Oriented Technology Analysis) [6]. Целью подобных исследований является разработка средств интеллектуальной поддержки систематического процесса обоснования возможных путей развития науки и технологий в различных областях, оценки перспективного влияния новых технологий на общество и окружающий мир, в том числе и на конкретные отраслевые инфраструктуры, а также поддержка выработки «скользящих» стратегических решений по инновационному развитию отраслей мировой экономики. Традиционно обоснованные научно-технологические прогнозы и программы инновационного развития являлись одной из важнейших устоявшихся форм регулирования экономики в таких странах, как США и Великобритания. Стремительное развитие информационных технологий и наступление «Эпохи Больших данных» (Big Data Age) вызвало значительный рост научных исследований в области технологического прогнозирования и в Китае [5, 7, 8]. Исследователи разрабатывают и эффективно применяют методы определения новых технологических решений (Tech Mining [9]), основанные на использовании интеллектуальных семантических технологий поиска, извлечения и анализа гетерогенных данных из электронных источников информации (Text Mining [10]) в соответствии с концепцией Big Data [11, 12].

Предлагаемый подход для семантического анализа Больших данных о научных и технологических решениях в области энергетики. Традиционные методы анализа данных о научных и технологических решениях (методы научно-технологического форсайта и системного анализа) и прогнозирование на их основе развития отрасли энергетики не всегда эффективны в силу отсутствия легкодоступной необходимой достоверной информации. При применении методов Big Data Analytics 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. В качестве источников информации для составления прогнозов можно использовать открытые данные (Linked Open Data), как из государственных информационных систем, так и 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 в автоматическом режиме. Кроме того, для поиска неструктурированной, но потенциально интересной для исследователей информации, могут использоваться Интернет-поисковые системы с предварительным анализом, классификацией и качественной 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. Найденные в различных информационных ресурсах документы должны быть идентифицированы, как потенциально полезные, классифицированы в соответствии с расширяемой моделью предметной области, которая может быть представлена в виде онтологии, а после отправлены в хранилище с дескриптивным описанием. Сканирование источников информации на регулярной основе позволяет не только обогащать хранилище знаний, но и отслеживать динамику изменения качественных и количественных показателей на основе когнитивных моделей.

Поиск информации и классификация на основе онтологии. Использование в качестве источников для анализа и последующего построения научно-технологических прогнозов развития энергетики исключительно русскоязычных информационных ресурсов является ошибочным, поскольку из-за глобализации инновационные разработки и результаты научных исследований не ограничены экономикой одной отдельно взятой страны или макрорегиона. Таким образом, возникает необходимость анализа информации не только на русском, но и на английском языке. Наличие множества публикаций на русском языке является одним из индикаторов готовности к практическому внедрению новой технологии в производственные процессы на предприятиях и организациях России. Решение задачи классификации документов и вычисления метрик выполняется на основе анализа словокомплексов [13, 14], при этом формально разными словокомплексами (в том числе и на разных языках) могут описываться фактически одинаковые понятия. Для решения указанной задачи было предложено применение билингвистической онтологии, включающей термины и определения (абстрактные и базовые концепты), а также поддерживающей синонимию (полную или частичную семантическую близость) в одном или нескольких различных языках. Пример структуры и связей концептов билингвистической онтологии приведен на рис. 1. Базовые концепты в онтологии могут быть идентифицированы по словокомплексам, которые в свою очередь, могут быть представлены устойчивыми словосочетаниями в русском или английском языке. Несколько базовых понятий (например, «бензин», «газ», «дизельное топливо») могут быть объединены в абстрактное понятие («топливо»). На рис. 2 представлен фрагмент билингвистической онтологии, определяющей понятия «древесное топливо/wood fuel», «уголь/coal», «биотопливо/biofuel» и «топливо/fuel».

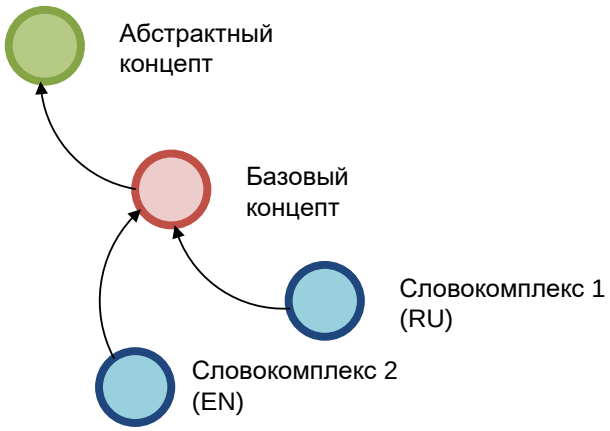

Рис. 1. Структура и связи концептов билингвистической онтологии для поддержки синонимии на нескольких языках.

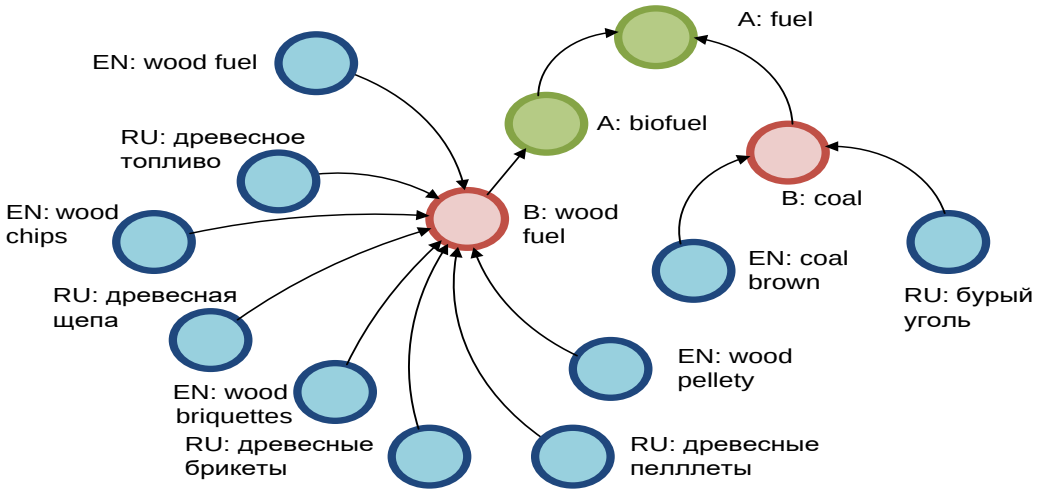

Рис. 2. Фрагмент билингвистической онтологии, определяющей абстрактные понятия (A), базовые понятия (B) и синонимы на нескольких языках (EN, RU).

С помощью онтологий можно описать иерархию энергетических технологий с достаточной детализацией их компонентов и взаимосвязей на разных уровнях (ресурсы, функции, типы преобразования энергии, потребительские услуги, инфраструктура, управление и т.д.); спецификации технологий или их характеристики с точки зрения технико-экономической эффективности; спецификации полного жизненного цикла энергетической технологии (Life Cycle Assessment); характеристики социально-экономических факторов; спецификации показателей развития инновационных технологий; определить концептуальную основу с двуязычными ссылками и синонимами для понятий всех уровней.

Поскольку анализ поступающей информации требует значительного времени ввиду большого исходного объема, авторами предлагается выделение двух процессов: процесса информационного наполнения хранилища и процесса использования наполненного хранилища исследователями для решения прикладных задач. Подсистема информационного наполнения состоит из модулей поиска 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 о научных и технологических решениях в области энергетики.

Представление метаданных в формате RIS (Research Information Systems). Подсистема информационного наполнения является точкой получения информации (метаданных, ссылок на внешние ресурсы и документов исследователей, описанных в понятиях системы онтологий), которая после извлечения может быть представлена пользователю или обработана другими программными агентами. 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 системами поддержки научных исследований и содержит описание отдельных ресурсов (как правило, научных публикаций) в разрезе до 79 параметров, основными из которых являются название (TI/T1/T2), информация об авторах (AU/AD), тип (TY) и дата публикации (PY/DA), база индекса (DB). Пример описания патента и описания статьи в журнале в формате RIS приведен на рис. 3.

а)

TY - PAT
CY - US
M3 - B2
SN - US 9391462 B2
ID - 180-107-644-067-313
C2 - 2016/07/12
PY - 2016
M1 - US 201213596618 A
DA - 2012/08/28
C1 - 2012/08/28
TI - Energy Storage System With Wired And Wireless Energy Transfer Function
AU - YANG YIL SUK
AU - KIM JONG DAE
AU - HEO SE WAN
AU - OH JI MIN
AU - KIM MIN KI
AU - KWON JONG KEE
PB - YANG YIL SUK
PB - KIM JONG DAE
PB - HEO SE WAN
PB - OH JI MIN
PB - KIM MIN KI
PB - KWON JONG KEE
PB - KOREA ELECTRONICS TELECOMM
OP - KR 20120008906 A 20120130
ER - *(конец документа)*

б)

TY - JOUR
TI - Application of ANFIS-PID controller for statcom to enhance power quality in power system connected wind energy system
T2 - International Journal of Engineering and Technology(UAE)
J2 - Int. J. Eng. Technol.
VL - 7
IS - 4
SP - 35
EP - 37
PY - 2018
DO - 10.14419/ijet.v7i4.4.19604
SN - 2227524X (ISSN)
AU - Huu Vinh, N.
AU - Hung, N.
AU - Kim Hung, L.
AD - Hochiminh City Power Company, Viet Nam
AD - Hochiminh City University of Technology (HUTECH), Viet Nam
AD - Danang University of Technology, Viet Nam
KW - Adaptive neuro-fuzzy inference system - proportional integral derivative (ANFIS-PID)
KW - Power quality
KW - Static synchronous compensator (STATCOM)
KW - Wind energy
PB - Science Publishing Corporation Inc
N1 - Export Date: 17 December 2019
M3 - Article
DB - Scopus
LA - English
UR - https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053468375&doi=10.14419%2fijet.v7i4.4.19604&partnerID=40&md5=e2d06aede670d0cdcb0cae5a3e6ce924
ER - *(конец документа)*

Рис. 3. Пример RIS-записи: а) патент, б) статья в журнале.

Для анализа информации, размещенной на различных сайтах в сети Интернет, на первом этапе необходимо получение ссылок на такие информационные ресурсы. Наиболее эффективным решением, с точки зрения авторов, является использование баз данных

распространенных и хорошо известных информационно-поисковых систем, выполнивших индексацию информационных ресурсов в сети Интернет. Наиболее распространенными в российском сегменте сети Интернет являются системы Яндекс (21,9%) и Google (68,8%), последняя также наиболее распространена и за пределами русскоговорящих стран (90,7%). Для дальнейшей унификации обработки информации извлеченные из них данные могут быть представлены в формате RIS и использоваться как самостоятельные источники информации, так и являться точками для начала углубленного анализа информационных ресурсов программами-краулерами (поисковыми роботами). Авторами было выполнено расширение RIS-формата и введен новый тип ресурсов (TY=IRES), при этом в качестве базы данных (DB) указывается наименование поисковой системы, а извлеченное описание ресурса помещается в поле Abstract (AB). Пример описания информационного ресурса, извлеченного из поисковой системы, приведен на рис. 4.

```
TY  - IRES
TI  - What is green energy? | MNN - Mother Nature Network
UR  - https://www.mnn.com/earth-matters/energy/stories/what-is-green-energy
AB  - Jul 25, 2012 - Green energy comes from natural sources such as sunlight, wind, rain, tides,
      plants, algae and geothermal heat. These energy resources are renewable, meaning they're
      naturally replenished. In contrast, fossil fuels are a finite resource that take millions of years
      to develop and will continue to diminish with use.
DB  - Google
ER
```

Рис. 4. Пример RIS-записи для поисковой системы.

Реализация инструментальных средств семантического анализа. При реализации интеллектуальной информационной системы используется сервис-ориентированный подход, позволяющий выполнять независимую разработку отдельных компонентов системы, что обеспечивает общую гибкость. На рис. 5 показана архитектура интеллектуальной информационной системы, включающая средства семантического анализа текстовых данных, онтологического, когнитивного и событийного моделирования, средства проверки гипотез и визуализации результатов поиска [15]. При обработке результатов поисковых запросов во внешних, по отношению к интегрированному хранилищу, программных средствах извлекаемых данных исследователи получают новые знания, которые могут быть представлены в явной форме и загружены в интегрированное хранилище с использованием средств коллективной работы.

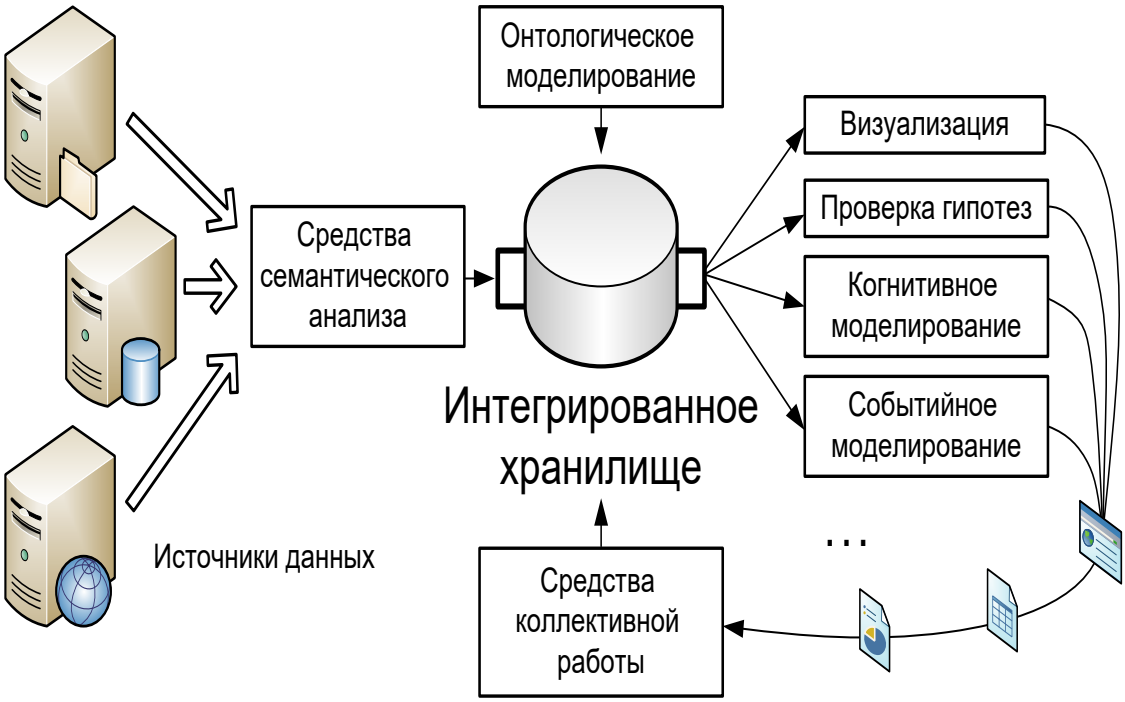

Рис. 5. Архитектура интеллектуальной информационной системы.

Применение для связи с внешними компонентами стандартизированных интерфейсов и сетевых протоколов Интернет, в частности JSON и HTTP, обеспечивает интеграцию методов при решении нетривиальных задач и позволяет получить доступ к информации и функциональным компонентам, размещенным на удаленных серверах.

При реализации средств сематического анализа используются как статистические методы, так и лингвистический анализ [16]. В статистических методах используется анализ частоты встречаемости слов (подсчет количества появлений слов в различных фрагментах, распределении частоты по документам и т.д.), при лингвистическом анализе выполняется идентификация отдельных слов, анализ их морфологических особенностей, выделение основы (стемминг), синтаксический и семантический анализ фрагментов текста. Технология стемминга играет важную роль в улучшении результатов анализа и сокращении области поиска, что позволяет идентифицировать основу слова, связывать многие формы одного и того же слова друг с другом и значительно упростить обработку текстовых массивов. Реализация средств семантического анализа выполнена на языке программирования Python с использованием библиотеки для лингвистического анализа текстов Natural Language Toolkit (NLTK) [17].

Первым этапом при проведении исследования является построение терминологического словаря предметной области, построение онтологии на его основе и расчет векторных представлений для концептов онтологии. При обработке массивов текстовых данных на первом этапе производится фильтрация и отделение от часто используемых слов и других подобных языковых элементов, которые не отражают специфику выбранной предметной области. После этого выполняются классификация и семантическое сравнение в зависимости от языка с элементами онтологий «RU» и «EN». Далее выявляются наиболее характерные признаки на основе расчета метрик TF-IDF для каждого поступающего документа. Применение показателя TF-IDF позволяет оценить важность слова или понятия в контексте документа. TF-IDF словокомплекса пропорционален частоте его использования в документе и обратно пропорционален частоте его использования во всех документах корпуса.

После размещения коллекции документов в интегрированном хранилище возможно проведение аналитического исследования и оценка гипотез о перспективах развития тех или иных новых технологий. При этом исследователь имеет возможность использовать только те источники, которым он доверяет. В общем случае гипотезы формируются на основе анализа массива собранных и отфильтрованных данных. Поскольку данные собираются из открытых источников, они явно носят статистический характер и, следовательно, к ним применимы статистические методы, в том числе методы статистической оценки гипотез и критериев. Ввиду большого объема собираемых данных целесообразно применение второго уровня статистической значимости (0,01 или 1%), что задает уровень статистической ошибки первого рода не более 1% и достоверность получаемых результатов при проверке гипотез не менее 99%.

Заключение. Реализация интеллектуальной информационной системы на основе сервис-ориентированного подхода и интеграции существующих сторонних и авторских компонентов позволяет эффективно решать задачи семантического анализа Больших данных о научных и технологических решениях в области энергетики. Средства семантического анализа разработаны на языке Python с использованием библиотеки NLTK. Настройка системы семантического анализа выполняется с использованием билингвистической онтологии, которая определяет абстрактные и базовые понятия, а также позволяет задавать синонимы и обеспечивает благодаря этому поддержку нескольких языков. Применение предложенных методов и подходов обеспечивает достаточный уровень гибкости и расширяемости.

Благодарности. Результаты получены в рамках проекта по госзаданию ИСЭМ СО РАН АААА-А21-121012090007-7 (проект № FWEU-2021-0007), при частичной финансовой поддержке грантов РФФИ №20-07-00994 и № 19-07-00351 с использованием ресурсов ЦКП "Высокотемпературный контур" (Минобрнауки России, проект № 13.ЦКП.21.0038).

СПИСОК ЛИТЕРАТУРЫ

1. Массель Л.В., Массель А.Г. Интеллектуальные вычисления в исследованиях направлений развития энергетики // Известия Томского политехнического университета. 2012. Т. 321. № 5. Управление, вычислительная техника и информатика. С. 135-141.
2. Массель Л.В. Создание и интеграция интеллектуальных информационных технологий и ресурсов для комплексных исследований в энергетике // Вестник РФФИ. 2012. № 4. С. 74-81.
3. Массель А.Г., Пяткова Н.И. Применение методов когнитивного моделирования для анализа угроз энергетической безопасности // Информационные и математические технологии в науке и управлении. 2020. № 4 (20). С. 24-33. DOI: 10.38028/ESI.2020.20.4.002.
4. Coates V., et al. On the future of technological forecasting / Technol. Forecast. Soc. Change. 67 (1). 2001. Pp. 1-17.
5. Zhang Y., et al. Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research // Technol. Forecast. Soc. Change. 2016. DOI: 10.1016/j.techfore.2016.01.015.
6. Future-Oriented Technology Analysis. Strategic Intelligence for an Innovative Economy / Eds. C. Cagnin et al. Springer. 2008. 170 p. DOI: 10.1007/978-3-540-68811-2.

7. Zhang M.L, Zhou Z.H. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization // IEEE Transactions on Knowledge and Data Engineering. Vol. 18. No. 10. 2006. Pp. 1338-1351. DOI: 10.1109/TKDE.2006.162.
8. Zheng L., Noroozi V., Yu P.S. Joint Deep Modeling of Users and Items using Reviews for Recommendation. 2017. DOI: 10.1145/3018661.3018665.
9. Cunningham, S.W., Porter, A.L., and Newman, N.C. Tech Mining, special issue of Technological Forecasting & Social Change. 73 (8). 2006. Pp. 915-1060.
10. Mirhosseini M. A Clustering Approach using a Combination of Gravitational Search Algorithm and k-Harmonic Means and its Application in Text Document Clustering // Turkish Journal of Electrical Engineering and Computer Sciences. Vol. 25. No. 2. 2017. Pp. 1251-1262. DOI: 10.3906/elk-1508-31.
11. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. М.: Манн, Иванов и Фербер. 2014. 230 с.
12. Марц Н., Уоррен Дж. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени. М.: Вильямс. 2016. 368 с. ISBN: 978-5-8459-2075-1.
13. Сидорова Е.А., Анохин С.В., Кононенко И.С., Саломатина Н.В. Тематический анализ запросов пользователей на основе предметно-ориентированного словаря // Вестник НГУ. Серия: Информационные технологии. 2014. №4. Режим доступа: https://cyberleninka.ru/article/n/tematicheskiy-analiz-zaprosov-polzovateley-na-osnove-predmetno-orientirovannogo-slovarya (дата обращения: 18.11.2021).
14. Dikovickij V.V., Shishaev M.G. Natural language text processing in search engine models // Proceedings of the Kola Scientific Center of the Russian Academy of Sciences. 3.
15. Копайгородский А.Н., Мамедов Т.Г. Архитектура интеллектуальной информационной системы для поддержки экспертных решений по стратегическому инновационному развитию энергетики // Информационные и математические технологии в науке и управлении. 2020. № 4 (20). С. 168-176. DOI: 10.38028/ESI.2020.20.4.015.
16. Kopaygorodsky A. Natural Language Processing for Forecasting Innovative Development of the Energy Infrastructure // E3S Web Conf. 209 03015 (2020). DOI: 10.1051/e3sconf/202020903015.
17. Bird S., Klein E., Loper E. Natural Language Processing with Python. O'Reilly Media Inc. 2009. ISBN 978-0-596-51649-9.

UDC 004.82:004.912:001.18:620.9

DESIGN AND DEVELOPMENT OF INSTRUMENTAL TOOLS FOR SEMANTIC ANALYSIS OF BIG DATA SCIENTIFIC AND TECHNOLOGICAL SOLUTIONS IN THE FIELD OF ENERGY

Alex N. Kopaygorodsky
Ph.D., Leading specialist, Department “Artificial Intelligence Systems in Energy”
e-mail: kopaygorodsky@isem.irk.ru
Elena P. Khairullina
Graduate student, Department “Artificial Intelligence Systems in Energy”
e-mail: lena-skoklenyova@yandex.ru
Melentiev Energy Systems Institute of SB RAS
664033 Irkutsk, Russia, Lermontov St., 130.

Abstract. The article discusses approaches to the design and implementation of individual components of instrumental tools for semantic analysis of information on scientific and technological solutions in the field of energy. This information has already been placed open sources. The structure of billinguistic ontology is considered, which makes it possible to solve the task of classifying information, taking into account its submission in various languages and synonyms. The authors reviewed the approach to the search and processing of information from open sources based on the use of semantic analysis developed by authors, the implementation of which was performed on Python using the Natural Language Toolkit library.

Keywords: Scientific and technological forecasting, semantic analysis, classification of text documents, billinguistic ontology.

Acknowledgments: The results were obtained within the framework of the project on the state assignment of the ISEM SB RAS AAAA-A21-121012090007-7 (project No. FWEU-2021-0007), with partial financial support from RFBR grants No. 20-07-00994 and No. 19-07-00351 using the resources of the Center for Collective Use "High-temperature circuit" (Ministry of Education and Science of Russia, project No. 13. TsKP.21.0038).

REFERENCES

1. Massel' L.V. Massel' A.G. Intellektual'nye vychislenija v issledovanijah napravlenij razvitija jenergetiki [Intellectual computing in the research of energy development directions] // Izvestia Tomsk Polytechnic University = Bulletin of the Tomsk Polytechnic University. 2012. Vol. 321. No. 5. Management, computing equipment and informatics. Pp. 135-141.
2. Massel' L.V. Sozdanie i integracija intellektual'nyh informacionnyh tehnologij i resursov dlja kompleksnyh issledovanij v jenergetike [Creation and integration of intelligent information technologies and resources for complex research in energy] // Vestnik RFFI = RFBR Bulletin. 2012. No. 4. Pp. 74-81.
3. Massel' A.G., Pjatkova N.I. Primenenie metodov kognitivnogo modelirovanija dlja analiza ugroz jenergeticheskoj bezopasnosti [Application of cognitive modeling methods for energy security threats analysis] // Informatsionnyye i matematicheskiye tekhnologii v nauke i upravlenii = Information and mathematical technologies in science and management. 2020. No. 4 (20). Pp. 24-33. DOI: 10.38028/ESI.2020.20.4.002.
4. Coates V., et al. On the future of technological forecasting / Technol. Forecast. Soc. Change. 67 (1). 2001. Pp. 1-17.

5. Zhang Y., et al. Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research // Technol. Forecast. Soc. Change. 2016. DOI: 10.1016/j.techfore.2016.01.015.
6. Future-Oriented Technology Analysis. Strategic Intelligence for an Innovative Economy / Eds. C. Cagnin et al. Springer. 2008. 170 p. DOI: 10.1007/978-3-540-68811-2.
7. Zhang M.L, Zhou Z.H. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization // IEEE Transactions on Knowledge and Data Engineering. Vol. 18. No. 10. 2006. Pp. 1338-1351. DOI: 10.1109/TKDE.2006.162.
8. Zheng L., Noroozi V., Yu P.S. Joint Deep Modeling of Users and Items using Reviews for Recommendation. 2017. DOI: 10.1145/3018661.3018665.
9. Cunningham S.W., Porter A.L., Newman N.C. Tech Mining, special issue of Technological Forecasting & Social Change. 73 (8). 2006. Pp. 915-1060.
10. Mirhosseini M. A Clustering Approach using a Combination of Gravitational Search Algorithm and k-Harmonic Means and its Application in Text Document Clustering // Turkish Journal of Electrical Engineering and Computer Sciences. Vol. 25. No. 2. 2017. Pp. 1251-1262. DOI: 10.3906/elk-1508-31.
11. Mayer-Shenberger V., Cukier K. Bol'shie dannye. Revoljucija, kotoraja izmenit to, kak my zhivem, rabotaem i myslim [Big Data. A Revolution That Will Transform How We Live, Work and Think]. Moscow: Mani, Ivanov and Ferber. 2014. 230 p.
12. Marz N., Warren J. Bol'shie dannye. Principy i praktika postroenija masshtabiruemyh sistem obrabotki dannyh v real'nom vremeni [Big Data: Principles and best practices of scalable realtime data systems]. Moscow: Wilyams. 2016. 368 p. ISBN: 978-5-8459-2075-1.
13. Sidorova E.A., Anohin S.V., Kononenko I.S., Salomatina N.V. Tematicheskij analiz zaprosov pol'zovatelej na osnove predmetno-orientirovannogo slovarja [Thematic analysis of user requests based on an object-oriented dictionary] // Bulletin NSU. Series: Information Technology. 2014. No. 4. Available at: https://cyberleninka.ru/article/n/tematicheskiy-analiz-zaprosov-polzovateley-na-osnove-predmetno-orientirovannogo-slovarya (accessed 18.11.2021).
14. Dikovickij V.V., Shishaev M.G. Natural language text processing in search engine models, Proceedings of the Kola Scientific Center of the Russian Academy of Sciences. 3.
15. Kopaygorodsky A.N., Mamedov T.G. Arhitektura intellektual'noj informacionnoj sistemy dlja podderzhki jekspertnyh reshenij po strategicheskomu innovacionnomu razvitiju jenergetiki [Architecture of the Intellectual Information System to support expert decisions on the strategic innovative energy development] // Informatsionnyye i matematicheskiye tekhnologii v nauke i upravlenii = Information and mathematical technologies in science and management. 2020. No. 4 (20). Pp. 168-176. DOI: 10.38028/ESI.2020.20.4.015.
16. Kopaygorodsky A. Natural Language Processing for Forecasting Innovative Development of the Energy Infrastructure // E3S Web Conf. 209 03015 (2020). DOI: 10.1051/e3sconf/202020903015.
17. Bird S., Klein E., Loper E. Natural Language Processing with Python. O'Reilly Media Inc. 2009. ISBN 978-0-596-51649-9.

Статья поступила в редакцию 15.12.2021; одобрена после рецензирования 22.12.2021; принята к публикации 24.12.2021.

The article was submitted 15.12.2021; approved after reviewing 22.12.2021; accepted for publication 24.12.2021.