1.3. ВИЗУАЛИЗАЦИЯ ПРОГНОЗА ТРЕНДОВ НАУЧНЫХ ТЕМ ДЛЯ ОПРЕДЕЛЕНИЯ ПЕРСПЕКТИВНЫХ НАПРАВЛЕНИЙ В ОБЛАСТИ БЕЗОПАСНОСТИ АЭС

Шарнин М.М. к.т.н., ФИЦ ИУ РАН, Москва,
Тищенко А.С., к.э.н., ЦЭНО ИПЭИ РАНХиГС, Москва,

В статье представлены результаты применения метода визуализации долгосрочного прогнозирования трендовых тем исследований в области обеспечения безопасности АЭС. Значимые темы были определены среди слов, включенных в названия научных статей. Формулировки заголовков, которые несколько раз встречаются в цитируемых статьях анализируемой коллекции, рассмотрены как трендовые темы исследования. Длительность роста тренда цитирования была целью для алгоритмов машинного обучения. Использовался метод машинного обучения CatBoost. Для визуализации прогноза были использованы методы t-SNE и Word2Vec. Кластеры трендовых ключевых слов на семантической карте помогли точно определить перспективные направления в области обеспечения безопасности АЭС.

Введение

Прогнозирование трендов стало чрезвычайно популярным во многих отраслях промышленности и в научной литературе. Долгосрочное прогнозирование трендовых тем исследований, основанное на анализе библиографических коллекций миллионов научных статей, помогает определить перспективные тенденции направления, найти прорывные идеи и сосредоточить усилия в наиболее продуктивном направлении. Поиск релевантных тем из набора публикаций рассмотрен в существующих исследованиях [Mei & Zhai, 2005; Jamali, H. R. & Nikzad, M., 2011].

В статье представлен метод визуализации долгосрочного прогнозирования трендовых тем исследований и результаты применения этого метода в области безопасности атомных электростанций (АЭС). Авторы разработали метод долгосрочного прогнозирования научных трендов на основе академических больших данных и системы специально подобранных параметров/характеристик ключевых слов в этих данных. Метод использует известные алгоритмы CatBoost, t-SNE и Word2Vec.

Разработанный авторами метод долгосрочного прогнозирования тенденций был апробирован на коллекции из 5 миллионов научных статей. В результате были определены ключевые слова, встречающиеся в статьях по теме безопасности АЭС и имеющие наиболее долгосрочные тенденции в будущем, и проведена их визуализация. Семантическое сходство между этими ключевыми словами было определено нейронной сетью Word2Vec и визуализировано на семантической карте. Далее были определены перспективные области, представленные кластерами ключевых слов. Некоторые из этих кластеров включают такие ключевые слова, как машинное обучение, нейронная сеть, аварии, человеческие ошибки и другие. Эти ключевые слова напрямую связаны с диагностикой безопасности АЭС, спасательными роботами и минимизацией риска человеческих ошибок. Эта связь также подтверждена примерами статей из сборника.

Визуальная аналитика позволяет быстро определить перспективные направления, представленные кластерами трендовых ключевых слов на семантической карте. Близкое расположение трендовых ключевых слов взаимно подтверждает точность прогноза их будущих тенденций. Слова, расположенные рядом с трендовыми ключевыми словами на семантической карте (например, диагностика, спасение), помогают понять специфику, связи и предназначение этих ключевых слов. Примерами кластера ключевых слов, которые объясняют друг друга, являются machine learning (машинное обучение), diagnostics (диагностика), neural network (нейронная сеть), wall-climbing (стенолазный), rescue robots (роботы-спасатели). Этот вид визуальной аналитики также подкрепляется отобранными примерами названий статей с этими ключевыми словами.

Разработанный авторами метод демонстрируется на примере анализа статей по теме безопасности АЭС. Данная работа представляет результаты этого анализа и может рассматриваться как обзор научных статей в области безопасности АЭС, указывающий на перспективные направления в этой области.

Тенденции ключевых слов изучаются через динамику различных показателей в группах статей, содержащих эти ключевые слова. Наиболее важным показателем тренда ключевого слова является количество цитирований ключевого слова (ЦКС). Продолжительность роста тренда (равная количеству лет непрерывного роста его среднего числа цитирований) являлась целью для алгоритма машинного обучения.

Таким образом, для каждого ключевого слова, включенного в заголовки статей в коллекции, был построен временной ряд из различных показателей/характеристик этого ключевого слова. На основе этих временных рядов и машинного обучения была построена регрессионная модель и прогноз продолжительности роста тренда. Продолжительность роста тренда – это скалярная функция, которую модели машинного обучения вычислили для каждого ключевого слова. При построении временного ряда и расчете ошибок прогноза использовался параметр максимальной продолжительности прогноза (1, 3, 5 и 7 лет), который уменьшал реальную и прогнозную продолжительность, если их значения превышали значение этого параметра.

Результаты экспериментов, оценивающих точность прогнозирования долгосрочного роста трендов на 1, 3, 5 и 7 лет, представлены в разделе «Эксперименты». Следует отметить, что модель машинного обучения позволяет оценить значимость каждого показателя/характеристики по его влиянию на точность

прогноза. Было отмечено, что этот эффект различается для разных долгосрочностей прогноза (1, 3, 5 и 7 лет), а для некоторых показателей их значимость возрастает с увеличением долгосрочности. Таким образом, анализ точности долгосрочного прогноза позволил найти значимые показатели/характеристики.

Вклад данной работы заключается в следующем:

- разработан новый алгоритм долгосрочного прогнозирования будущих тенденций в научных статьях и определена точность прогноза;
- впервые создан прогнозный обзор научных статей в области безопасности АЭС, описывающий долгосрочные будущие тенденции с известной точностью прогноза;
- предложена новая форма визуализации долгосрочного прогноза будущих тенденций в научной литературе.

Обзор литературы

Одной из важных тем исследований является предсказание тенденций развития науки. Ниже рассмотрены релевантные работы с фокусом на точности долгосрочных предсказаний тематических трендов при наличии соответствующего содержания в работах.

В своем обзоре [Hou et al, 2019] Хоуа и другие проанализировали методы и приложения в области прогнозирования на основе данных в наукометрии и обсудили их значение. Они обобщили проблемы исследования с трех точек зрения: предсказание влияния статей, предсказание влияния ученых и предсказание сотрудничества авторов. Авторы не затронули проблему прогнозирования тематических трендов, что скорее всего связано с относительно небольшим количеством статей на эту тему.

Вопросам обнаружения тем и предсказания будущих трендов посвящена работа [Hurtado et al, 2015]. В этой работе авторы предлагают, используя ассоциативный анализ и ансамблевое прогнозирование, автоматически выявлять темы из набора текстовых документов и прогнозировать тенденции их развития в ближайшем будущем. Авторы также отмечают, что в области прогнозирования тем на чувствительных ко времени ресурсах, таких как Twitter, был достигнут значительный прогресс, но в научных публикациях существует очень мало работ.

Работа [Prabhakaran et al, 2016] посвящена прогнозированию подъема и спада научных тем на основе тенденций в их риторическом обрамлении. Авторы отмечают, что мало известно о механизмах, лежащих в основе роста и падения тем. Авторы обнаружили, что риторическая функция темы в значительной степени предсказывает ее окончательный рост или упадок. Например, темы, которые риторически описываются как результаты, имеют тенденцию к упадку, в то время как темы, которые функционируют как методы, как правило, находятся на ранних стадиях роста.

Работа [Shen, et al, 2016] посвящена моделированию академического влияния на уровне темы в научной литературе. В этой работе авторы вводят J-Index, количественную метрику для моделирования академического влияния статьи. Для каждой статьи J-Index учитывает количество ее цитирований, силу каждого цитирования и новизну всех статей, в которых она цитируется. Авторы предлагают модель ссылочной темы (RefTM) для измерения новизны каждой статьи, а также силы цитирования среди них. RefTM может эффективно обнаруживать темы высокого качества, моделировать новизну статьи и прогнозировать цитируемость.

Для прогнозирования используются различные методы машинного обучения. Деревья решений и нейросетевые методы демонстрируют хорошее качество прогнозирования и классификации, по сравнению с такими методами машинного обучения, как Random Forest, Support Vector Machine, Naive Bayes и K Nearest Neighbor. В работе [Karakurt, et al, 2016] сравнивались ансамбли деревьев решений и нейронных сетей для прогнозирования параметров на один день вперед. Результаты, полученные в этом исследовании, показывают, что ансамблевые модели обучения дают лучшую точность прогнозирования, чем обычная модель искусственной нейронной сети. Более того, ансамбли искусственных нейронных сетей превосходят ансамбли на основе деревьев.

В данной работе мы использовали метод машинного обучения CatBoost, который является одной из лучших реализаций деревьев решений [Prokhorenkova et al, 2019]. CatBoost – это высокопроизводительная библиотека с открытым исходным кодом для градиентного бустинга на деревьях решений. CatBoost - это реализация упорядоченного бустинга, перестановочной альтернативы классического алгоритма, которая была создана для борьбы со сдвигом предсказания, вызванным особым видом потерь целевых ориентиров, присутствующим во всех существующих на данный момент реализациях алгоритмов градиентного бустинга.

Коллекция DBLP

В наших экспериментах мы анализировали библиографию по информатике под названием DBLP citation network, которая представляет собой коллекцию статей с 1936 по 2020 год, составленную на ресурсе aminer.org и упоминаемую здесь как коллекция DBLP. Библиография по информатике DBLP предоставляет открытую библиографическую информацию по основным журналам и публикациям по информатике.

Данные о цитировании извлекаются из DBLP (Digital Bibliography & Library Project, dblp.org), ACM (Association for Computing Machinery, acm.org), MAG (Microsoft Academic Graph) и других источников. Для аналитической обработки использована версия V12, выпущенная в апреле 2020 года. Этот набор данных состоит из 4 894 081 статьи и 45 564 149 связей цитирования. При проведении аналитической обработки данных учтены все названия, места публикаций, авторы и связи цитирования. В данной работе коллекция DBLP была проанализирована в различных направлениях, описанных в следующем разделе.

Эксперименты

В данной работе изучены тренды ключевых слов через динамику различных показателей/характеристик групп статей, содержащих эти ключевые слова. Наиболее важной характеристикой трендового ключевого слова является показатель цитируемости ключевого слова (ЦКС). Для каждого ключевого слова рассчитан долгосрочный рост его тренда, который равен количеству лет непрерывного роста скользящего среднего значения тренда ЦКС. Эта продолжительность роста тренда была целевой для алгоритма машинного обучения CatBoost.

Регрессионная модель CatBoost была обучена для 20 признаков тренда слова/ключевого слова, включая: количество лет роста будущего тренда (предсказываемая); текущий ЦКС (количество признаков за последние 6 лет); общее количество статей со словом; количество лет роста предыдущего тренда; общий рост цитирования для предыдущего тренда; количество лет от начала тренда слова; количество ссылок цитирования между статьями с данным словом с начала тренда (в текущем и предыдущем годах – ряд признаков); количество лет от ситуации тренда до текущего/базового года T.

Таким образом, для каждого ключевого слова, включенного в заголовки статей в коллекции DBLP, были построены временные ряды из 20 различных показателей/характеристик этого ключевого слова. На основе этих временных рядов и метода машинного обучения были построены регрессионная модель и прогноз продолжительности роста тренда ЦКС. Продолжительность роста тренда ЦКС измерялась как количество лет роста скользящего среднего тренда ЦКС. Модель искала первый случай нисходящей тенденции у скользящего среднего и сообщала количество лет роста до этого момента. Скользящее среднее значение тренда ЦКС всегда рассчитывалось на 3 года, и это отличается от параметра D максимальной долгосрочности прогноза.

При составлении временных рядов и расчете ошибок прогнозирования использовался параметр D максимальной долгосрочности прогноза (1, 3, 5 и 7 лет), который ограничивает реальную и прогнозируемую продолжительность роста значением D. Параметр T также использовался для указания текущего/базового года. При обучении модели вся информация, относящаяся ко времени после T, была удалена. Точность и ошибка прогноза обученной модели проверялась в момент времени T + D на совершенно других данных, чем при обучении.

На рисунке 1 вертикальная ось представляет ошибку прогноза, а горизонтальная ось – текущий/базовый год эксперимента.

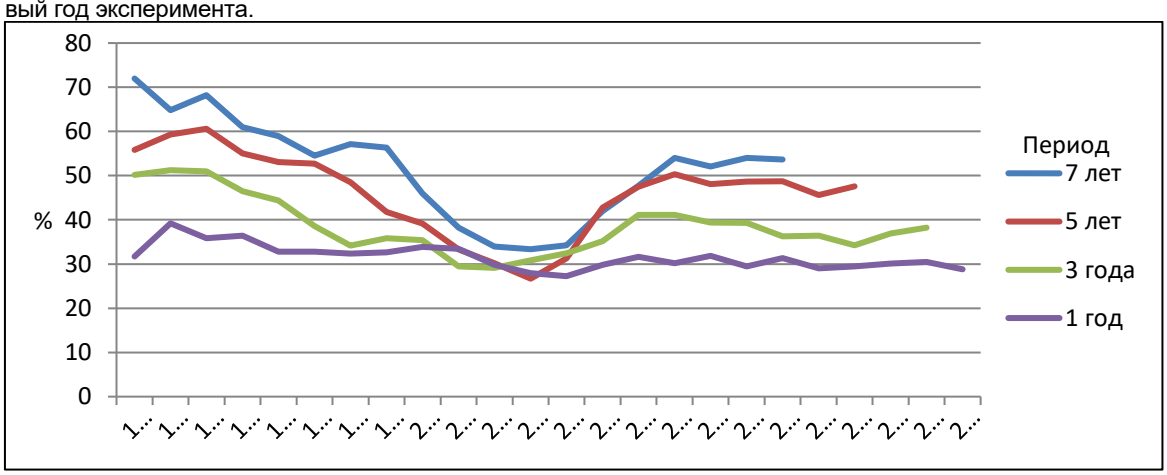

Рисунок 1. Ошибки прогнозирования долгосрочного роста трендов по годам (на 1, 3, 5 и 7 лет).

Видно, что, чем долгосрочнее прогноз, тем выше коэффициент ошибок. Уровень ошибок трехлетних прогнозов для ряда экспериментов с 1997 по 2014 год составил около 40%, в то время как уровень точности – 60%. Максимальная ошибка семилетних прогнозов для ряда экспериментов с 1997 по 2010 год составила около 58%. Ошибка семилетних прогнозов с 2000 по 2004 год составила менее 40%. На рисунке 1 показано количество ключевых слов с долгосрочными тенденциями на 1, 3, 5 и 7 лет в разные базовые годы. По вертикали обозначено количество ключевых слов с долгосрочными тенденциями, а по горизонтальной оси – текущий/базовый год эксперимента. Тренд считается долгосрочным, если его продолжительность превышает параметр D. Чем больше долгосрочных тенденций учитывается, тем выше точность.

В первые годы существования коллекции DBLP было меньше статей, меньше слов и, соответственно, меньше растущих трендов. В последние годы количество трендов уменьшалось, так как данных еще не было.

В результате прогностического эксперимента, при использовании только данных до 2018 года, было выявлено 5587 ключевых слов с положительной продолжительностью роста тренда. Наиболее выраженный долгосрочный прогноз продолжительности роста тренда (более 10 лет) имели следующие темы и их репрезентативные ключевые слова: artificial intelligence, AI, convolutional networks, CNN, CNN-based, deep network, using deep, via deep, explainable, unsupervised learning, learning-based, comprehension, reading comprehension, adversarial, adversarial learning, lstm, and local differential.

Таким образом, темы, связанные с искусственным интеллектом и нейронными сетями (CNN, LSTM), имели наиболее выраженные долгосрочные прогнозируемые тенденции. Этот автоматический прогноз совпадает с ожиданиями экспертов.

При исследовании темы БЕЗОПАСНОСТЬ АТОМНЫХ ЭЛЕКТРОСТАНЦИЙ был задан запрос: nuclear power (атомная энергетика) OR nuclear safety* (ядерная безопасность) OR nuclear PSA (nuclear probabilistic safety assessment, вероятностная оценка ядерной безопасности) OR nuclear decommis* вывод из эксплуатации) OR nuclear robot* (ядерный робот). Слова со звездочкой - это усеченные слова. На основании этого запроса из коллекции DBLP было отобрано 664 названия статей. Примеры отобранных названий приведены в разделе «Авторефераты».

Результаты исследования БЕЗОПАСНОСТЬ АТОМНЫХ ЭЛЕКТРОСТАНЦИЙ показали, что наиболее продолжительные тенденции роста имеют подтемы с ключевыми словами: accident (авария), preparedness (готовность), human error (человеческая ошибка), safety risk (риск безопасности), plants (заводы, станции, установки), stations (станции, терминалы), rescue robot (робот-спасатель), operator's (операторский), attention (внимание), safety-critical systems (критически важные для безопасности системы), cyber (кибернетический, виртуальный), transients (переходные процессы), develop (разработка, развитие), neural networks (нейронные сети), machine learning (машинное обучение). Этот автоматический прогноз согласуется с ожиданиями экспертов.

Визуализация

Для визуализации прогноза были использованы методы Word2Vec [Mikolov et al, 2013] и t-SNE [Van der Maaten L.J.P., Hinton G.E. , 2008]. Алгоритм визуализации состоит из следующих шагов: (1) анализируется коллекция научных статей, находятся трендовые ключевые слова и рассчитывается прогноз времени роста их трендов с помощью метода, описанного в разделе «Эксперименты». (2) С помощью метода Word2Vec рассчитывается семантическое сходство между трендовыми ключевыми словами и строится матрица сходства для трендовых ключевых слов. (3) Координаты точек на плоскости вычисляются с помощью метода t-SNE. (4) Строится визуальная карта, и ключевые слова отмечаются на ней точками разного цвета в зависимости от их прогноза.

Алгоритм t-SNE строит несколько различных семантических карт для одной матрицы сходства. На рисунке 2 показана семантическая карта, построенная этим методом для темы «БЕЗОПАСНОСТЬ АТОМНЫХ ЭЛЕКТРОСТАНЦИЙ».

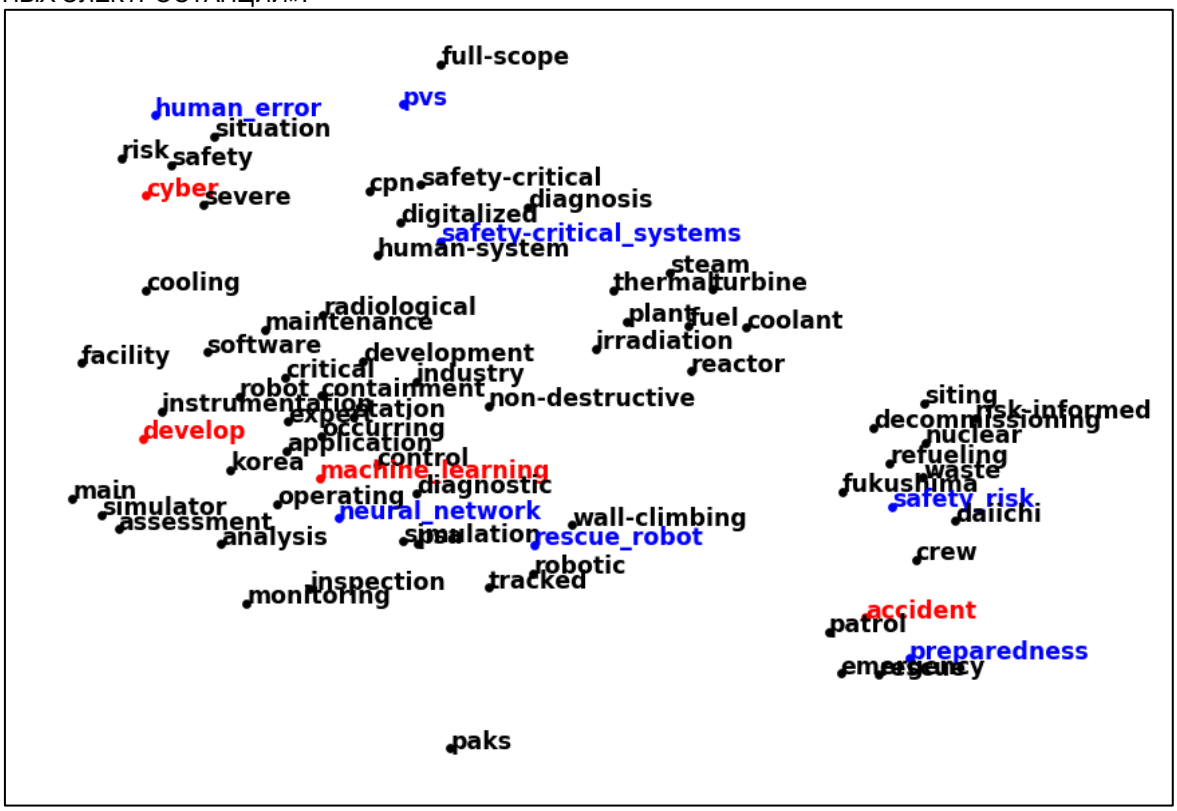

Рисунок 2. Визуализация ключевых слов по безопасности АЭС с долгосрочными трендами на основе t-SNE проекций матрицы сходства Word2Vec между ключевыми словами и прогнозом периода роста их трендов (красный: ключевые слова с самыми длинными трендами, синий: ключевые слова со средними трендами, черный: ключевые слова с самыми короткими трендами).

Семантически схожие ключевые слова группируются в кластеры на семантической карте. Как оказалось, эти кластеры в основном состоят из ключевых слов с одинаковой продолжительностью тренда. Таким образом, трендовые ключевые слова на семантической карте могут подтвердить длительность трендов друг друга. Пространственная близость семантически схожих ключевых слов обеспечивается с помощью технологии нейронных сетей, методов Word2Vec и t-SNE. Прогноз темы определяется как среднее значение прогнозов для ключевых слов в кластере, что повышает точность прогноза. Среднее значение используется для получения обобщающей характеристики некоторых наборов данных. Если данные более или менее однородны и нет аномальных наблюдений (выбросов), то среднее хорошо обобщает данные, что минимизирует влияние случайных факторов (они нивелируют друг друга при агрегации). Таким образом, трендовые ключевые слова на семантической карте подтверждают друг друга и помогают более точно определить перспективные направления.

Слова, расположенные рядом с трендовыми ключевыми словами на семантической карте (например, диагностика, спасение), помогают понять специфику, связи и предназначение этих ключевых слов. Примерами кластера ключевых слов, которые объясняют друг друга, являются являются machine learning (машинное обучение), diagnostics (диагностика), neural network (нейронная сеть), wall-climbing (стенолазный), rescue robots (роботы-спасатели). В следующем разделе этот вид визуальной аналитики также подкрепляется избранными примерами названий статей с этими ключевыми словами.

Авторефераты

Для объяснения трендовых тем, их прогнозирования и визуализации был автоматически составлен реферат из заголовков статей, содержащих трендовые ключевые слова. Эти ключевые слова выделены в тексте. Ниже приводится часть этих авторефератов (на английском и русском языках):

«Super-TOPIC: NUCLEAR POWER PLANT SAFETY

TRENDS: accident, preparedness, human_error, safety_risk,
plants, stations, rescue_robot,
operator's, attention, safety-critical_systems,
cyber, transients, develop,
neural_networks, neural_network, machine_learning.

TOPIC: ACCIDENT, PREPAREDNESS, SAFETY_RISK, EMERGENCY, HUMAN_ERROR, WASTE

Emergency response to the nuclear accident at the Fukushima Daiichi. Nuclear Power Plants using mobile rescue robots.

Development of a robotic system for nuclear facility emergency preparedness -- observing and work-assisting robot system.

Development of a dynamical systems model of plant programmatic performance on nuclear power plant safety risk.

Human error probabilities from operational experience of German nuclear power plants.

TOPIC: NEURAL_NETWORK, ANALYSIS, MACHINE_LEARNING

A neural networks design methodology for detecting loss of coolant accidents in nuclear power plants.

A convolutional neural network model for abnormality diagnosis in a nuclear power plant.

Machine learning of fire hazard model simulations for use in probabilistic safety assessments at nuclear power plants.

TOPIC: TRACKED, WALL-CLIMBING, ROBOTIC, RESCUE_ROBOT, ROBOT

Design and control of a tracked robot for search and rescue in nuclear power plant.

Wall-climbing robot for inspection in nuclear power plants.

Improvements to the rescue robot quince toward future indoor surveillance missions in the Fukushima Daiichi nuclear power plant.

. . .

МАКРО-ТЕМА: БЕЗОПАСНОСТЬ АТОМНЫХ ЭЛЕКТРОСТАНЦИЙ

ТРЕНДЫ: авария, готовность, человеческий_террор, безопасность_риск,
станции, спасательный_робот, оператор, внимание, критические_системы_безопасности,
кибер, переходные процессы, развиваться, нейронные_сети, машинное_обучение.

ТЕМА (РУБРИКА): АВАРИЯ, ГОТОВНОСТЬ, РИСК_БЕЗОПАСНОСТИ, ЧРЕЗВЫЧАЙНЫЕ СИТУАЦИИ, ЧЕЛОВЕЧЕСКИЕ_ОШИБКИ, ОТХОДЫ

Аварийное реагирование на ядерную аварию на АЭС "Фукусима-1". Атомные электростанции с использованием мобильных роботов-спасателей.

Разработка роботизированной системы для аварийной готовности ядерного объекта - наблюдающая и помогающая в работе роботизированная система.

Разработка динамической системной модели влияния программной деятельности станции на риск безопасности атомной электростанции.

Вероятность человеческих ошибок из опыта эксплуатации немецких атомных электростанций.

ТЕМА (РУБРИКА): НЕЙРОННАЯ_СЕТЬ, АНАЛИЗ, МАШИННОЕ_ОБУЧЕНИЕ

Методология проектирования нейронных сетей для обнаружения аварий с потерей теплоносителя на атомных электростанциях.

Модель сверточной нейронной сети для диагностики аномалий на атомной электростанции.

Машинное обучение имитационных моделей пожарной опасности для использования в вероятностных оценках безопасности на атомных электростанциях.

ТЕМА (РУБРИКА): ГУСЕНИЧНЫЙ, СТЕНОЛАЗНЫЙ, РОБОТ-СПАСАТЕЛЬ, РОБОТ

Проектирование и управление гусеничным роботом для поиска и спасения на атомной электростанции.

Стенолазный робот для инспекции на атомных электростанциях.

Усовершенствования робота-спасателя quince для будущих миссий по наблюдению в помещениях на атомной электростанции Фукусима Дайичи (Fukushima Daiichi)

. . . ».

Этот реферативный текст, состоящий из названий статей, помогает понять взаимосвязь трендовых ключевых слов и их тем.

Заключение

Впервые изучена проблема долгосрочного прогнозирования трендовых тем исследований на основе академических больших данных. Визуализация таких долгосрочных прогнозов помогает эффективно ориентироваться и оценивать темы исследований, определять перспективные направления и фокусировать усилия на них.

Визуализации построены в виде семантических карт. Визуализации позволили объединить ключевые слова в кластеры и темы, а также создать более точные предсказания для тем. Для пояснения трендовых тем был автоматически построен текст (реферат) из названий статей, содержащих трендовые ключевые слова.

Для демонстрации эффективности предложенной модели были проведены эксперименты на наборе научных данных, включающем миллионы публикаций. Коэффициент ошибки трехлетних прогнозов трендовых исследовательских тем в ряде экспериментов с 1997 по 2014 год составил около 40%, а коэффициент точности - 60%. Результаты прогностического эксперимента по теме БЕЗОПАСНОСТЬ АЭС, показали, что наиболее выраженные долгорастущие тренды имеют подтемы, связанные с ключевыми словами: accident (авария), preparedness (готовность), human error (человеческая ошибка), safety risk (риск безопасности), plants (заводы, станции, установки), stations (станции), rescue robot (робот-спасатель), operator's (операторский), attention (внимание), safety-critical systems (критически важные для безопасности системы), cyber (кибернетический, виртуальный), transients (переходные процессы), develop (разработка, развитие), neural networks (нейронные сети), machine learning (машинное обучение).

Визуализации, построенные в виде семантических карт, позволили объединить ключевые слова в кластеры и темы, а также создать более точные прогнозы для тем. Для объяснения трендовых тем из названий статей, содержащих трендовые ключевые слова, был автоматически построен реферат.

Литература:

1. Mei Q, Zhai C. (2005) Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining. 2005.
2. Jamali, H. R., and Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. Scientometrics 88(2):653–661.
3. Hou, J. Pan H., Teng Guo T., Lee I., Kong X., Xia F. ()2019) “Prediction Methods and Applications in the Science of Science: A Survey”, Computer Science Review, Volume 34, November 2019, 100197, DOI 10.1016/j.cosrev.2019.100197
4. J. Hurtado, S. Huang and X. Zhu, "Topic Discovery and Future Trend Prediction Using Association Analysis and Ensemble Forecasting," 2015 IEEE International Conference on Information Reuse and Integration, 2015, pp. 203-206, doi: 10.1109/IRI.2015.40.
5. Prabhakaran, V.; Hamilton,W. L.; McFarland, D.; and Jurafsky, D. (2016). Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, 1170–1180.
6. Shen, J.; Song, Z.; Li, S.; Tan, Z.; Mao, Y.; Fu, L.; Song, L.; and Wang, X. (2016). Modeling topic-level academic influence in scientific literatures. In AAAI Workshop: Scholarly Big Data.
7. Karakurt, O., Erdal, H.I., Namli, E., Yumurtaci Aydogmus, H. and Turkan, Y.S. (2013), "Comparing ensembles of decision trees and neural networks for one-day-ahead streamflow prediction", Sci. Res. J.

8. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. (2019) CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516.
9. Mikolov T., Chen K., Corrado G., Dean J. (2013) Efficient Estimation of Word Representations in Vector Space // In Proceedings of Workshop at ICLR (2013).
10. Van der Maaten L.J.P., Hinton G.E. (2008) Visualizing Data Using t-SNE // Journal of Machine Learning Research. 2008, November (v. 9).

Михаил Михайлович Шарнин, к.т.н., старший научный сотрудник
ФИЦ ИУ РАН, Москва, Россия ORCID - 0000-0003-0450-5156
e-mail: 1@keywen.com

Michael Charnine, PhD, Senior Researcher, Federal Research Center "Informatics and Control", Russian Academy of Sciences,

Алексей Сергеевич Тищенко, к.э.н., старший научный сотрудник
Институт прикладных экономических исследований,
Центр экономики непрерывного образования РАНХиГС, Москва, Россия
ORCID - 0000-0002-5834-5760
e-mail: mc@keywen.com

Alexey Tishchenko, Senior Researcher
Center of the Economics for Continuing Education, Institute of Applied Economic Research, Russian Presidential Academy of National Economy and Public Administration (Moscow, Russia)
PhD in economics

Ключевые слова
Безопасность АЭС, визуализация, долгосрочное прогнозирование, трендовые темы исследований, дерево решений, CatBoost, научные статьи, динамика трендов тем, большие данных

Michael Charnine, Alexey Tishchenko, Visualization of the forecast of trends in scientific topics to identify promising areas in the field of NPP safety

Keywords
NPP safety, Visualization, Long-term forecasting, Trending research topics, Decision Tree, Cart Boost, scientific articles, topic trend dynamics, Big data.

DOI: 10.34706/DE-2022-03-03
JEL classification: C02 – Математические методы,

Abstract
The article presents the results of the application of the visualization method of long-term forecasting of trending research topics in the field of NPP safety. Significant topics were identified among the words included in the titles of scientific articles. The headline formulations that occur several times in the cited articles of the analyzed collection are considered as trending research topics. The duration of the citation trend growth was the goal for machine learning algorithms. The CatBoost machine learning method was used. tSNE and Word2Vec methods were used to visualize the forecast. Clusters of trending keywords on the semantic map helped to accurately identify promising areas in the field of nuclear power plant safety.