МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ИНФОРМАТИКА И КИБЕРНЕТИКА

2 (32)

Донецк – 2023

УДК 004.3+004.9+004.2+51.7+519.6+519.7

ИНФОРМАТИКА И КИБЕРНЕТИКА, № 2 (32), 2023,
Донецк, ДонНТУ.

Выпуск подготовлен по материалам XIV Международной научно-технической конференции «Информатика, управляющие системы, математическое и компьютерное моделирование – 2023» (ИУСМКМ–2023), проведенной 24 – 25 мая 2023 г. в рамках IX Международного Научного форума Донецкой Народной Республики. Представлены результаты научно-технической деятельности аспирантов, соискателей и научных работников. Статьи посвящены вопросам приоритетных направлений научно-технического обеспечения в области информатики, кибернетики, вычислительной техники и инженерного образования.

Материалы предназначены для специалистов народного хозяйства, ученых, преподавателей, аспирантов и студентов высших учебных заведений.

Редакционная коллегия
Главный редактор: *Павлыш В. Н., д.т.н., проф.*
Зам. глав. ред.: *Мальчева Р. В., к.т.н., доц.*
Ответственный секретарь: *Лёвкина А. И.*
Члены редакционной коллегии: *Аверин Г. В., д.т.н., проф.; Аноприенко А. Я., к.т.н., проф.; Звягинцева А.В., д.т.н., доц.; Зори С. А., д.т.н., доц.; Карабчевский В. В., к.т.н., доц.; Привалов М. В., к.т.н., доц.; Скобцов Ю. А., д.т.н., проф.; Федяев О. И., к.т.н., доц.; Шелепов В. Ю., д.ф-м.н., проф.*

Рекомендовано к печати ученым советом ФГБОУ ВО «Донецкий национальный технический университет» Министерства науки и высшего образования РФ. Протокол № 5 от 23 июня 2023 г.

Свидетельство о регистрации СМИ: серия ААА № 000145 от 20.06.2017.

Приказ МОН ДНР № 135 от 01.02.2019 о включении в Перечень рецензируемых научных изданий ВАК ДНР.

Контактный адрес редакции
ДНР, 83001, г. Донецк, ул. Артема, 58, ФГБОУ ВО «ДонНТУ»,
4-й учебный корпус, к. 36., ул. Кобозева, 17.
Тел.: +7 (856) 301-07-35, +7 (949) 334-89-11
Эл. почта: infcyb.donntu@yandex.ru
Интернет: http://infcyb.donntu.ru

© Донецкий национальный технический университет
Министерство науки и высшего образования РФ, 2023

СОДЕРЖАНИЕ

Информатика и вычислительная техника

УДК 004.048, 004.622

Формирование датасета для решения задач машинного обучения

В. О. Вовченко, В. А. Светличная, Н. К. Андриевская
Донецкий национальный технический университет, г. Донецк
wizziglod@gmail.com, svvictoria@mail.ru, nataandr@yandex.ru

Аннотация

Статья посвящена описанию основных этапов формирования корпуса данных для машинного обучения, а также методов предобработки текстов. Приведены варианты решения таких проблем, как неполнота данных, очистка и преобразование данных. Выполнено кодирование категориальных данных. С помощью методов предобработки NLP подготовлен набор данных, который будет в дальнейшем использован при векторизации и решении задачи классификации методами машинного обучения.

Введение

Одной из постоянных задач менеджера отдела продаж любого предприятия по производству продукции, в том числе и косметической, является ежедневная работа с рекламациями. В процессе анализа поступающих рекламаций возникла необходимость разработки современного интеллектуального инструмента, применение которого автоматизировало бы процесс обработки рекламаций [1]. Повышение эффективности претензионной деятельности должно проводиться за счёт применения актуальных средств информационных технологий, в частности машинного обучения (МО) и интеллектуального анализа данных (ИАД) [2].

Интеллектуальная обработка документов позволяет преобразовать неструктурированные и полуструктурированные данные в удобный структурированный формат. Другими словами данные из документов (сообщения, PDF-файлы, сканы, электронные письма и т. д.) извлекаются и преобразуются в текстовые оцифрованные данные, готовые к обработке [3].

Решение задач, связанных с извлечением информации из текстов, часто требует наличия специально подготовленных текстовых коллекций, собранных, обработанных и размеченных в соответствии со спецификой решаемой задачи. Такие коллекции называются текстовыми корпусами. Корпус текстов — это вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то фрагменты текстов данной проблемной области и являются исходными материалами для дальнейшей обработки и получении датасетов, используемых в дальнейшем при МО и ИАД.

Датасет (англ. dataset) – это обработанный и структурированный массив данных, который состоит из двух основных компонентов: непосредственно объекта и его параметров. Параметры объекта обычно задают не словами, а цифрами [4].

Информация в датасете представляет собой размеченные данные, которые являются основой для машинного обучения. Датасеты бывают различных видов:

1. Таблица, в строках которой расположены данные, а в колонках – параметры.
2. Граф, содержащий данные о связях между объектами. Данные графовой модели можно представить в виде таблицы, где в строках и колонках указаны данные, а в пересечениях – связи между ними.
3. Упорядоченные записи для данных, для которых основную роль играет конкретное расположение в таблице.

Часто бывает так, что датасетов по конкретному запросу не существует. В таком случае датасет приходится формировать самостоятельно.

Целью данной статьи является исследование этапов и методов предобработки текстов для формирования собственного датасета для МО.

Этапы предподготовки данных

Основная часть материалов по МО и ИАД посвящена описанию самих методов и алгоритмов, методам же предобработки данных и формированию корпусов данных уделено чрезвычайно мало внимания. Модели и алгоритмы машинного обучения описаны с точки зрения их применения на чистых, уже подготовленных данных. При этом, практикам МО и ИАД, хорошо известно, насколько значим вклад данных, а точнее уровень их подготовки перед использованием алгоритмами машинного обучения, в успешном решении поставленных задач. Место процесса предобработки данных в типовом процессе решения задач с использованием МО изображены на рисунке 1.

Рисунок 1 – Типовой процесс решения задач с МО и ИАД

Целью первого этапа сбора и анализа данных является обеспечение ясности данных и оценка полноты данных. Для этого надо оценить, достаточно ли количество имеющихся данных для решения задачи, насколько полно имеющиеся переменные описывают исследуемый процесс, и какие внешние факторы могут оказывать влияние на исследуемый процесс.

При решении задач первого этапа следует активно сотрудничать с экспертами в предметной области. Завершается этап составлением описания технологического набора данных, по возможности с формализацией и визуализацией данных.

Цель второго этапа – обеспечить полноту, корректность, непротиворечивость данных. Этап может включать в себя процедуры заполнения пропусков или восстановления данных, поиск невозможных значений и дубликатов, исправление форматов и сглаживание выбросов.

Методов обработки пропусков числовых данных предостаточно. Для обработки текстовых данных целесообразно применять следующие методы:

- если данных достаточно много, удалить строки с пропусками, с выбросами, с дубликатами;
- если данных ограниченное количество, заполнить самым частым значением или константой; заполнить случайными значениями выборки из аналогичного распределения.

В случае рассматриваемой задачи использовался метод заполнения случайными значениями из существующей выборки.

Цель третьего этапа заключается в том, чтобы обеспечить структурированность, однородность и согласованность данных.

Этап может включать в себя следующие процедуры: приведение типов данных и кодирование номинативных переменных, нормализацию данных, стандартизацию данных, обогащение данных, оптимизацию пространства признаков.

Задача процесса нормализации – улучшить качество работы алгоритмов за счёт приведения данных к нужному диапазону. К основным методам можно отнести: нормализацию на максимум; нормализацию на интервале; ранговую нормализацию.

Задача стандартизации заключается в улучшении качества работы алгоритмов за счёт приведения данных к стандартному нормальному распределению.

Нормализация и стандартизация существенно повышают эффективность метрических алгоритмов классификации: метода ближайшего соседа (k-Nearest Neighbors), метода k-средних (k-means), метода машин опорных векторов (Support Vector Machine).

После завершения всех этапов предварительной обработки данных выполняется векторизация текстов, так как алгоритмы машинного обучения предназначены для работы с числовыми данными, и необходимо выполнить преобразование текста отзыва в числовой вектор признаков.

Формирование корпуса

Поскольку отдел продаж разрабатывает систему учета рекламаций и отзывов сравнительно недавно, то имеющегося в наличии количества рекламаций является явно недостаточно для полноценного обучения нейронной сети. Возникла проблема неполноты данных.

Для ее решения были реализованы следующие подходы:

1. Способ получения данных из отзывов на продукцию данного предприятия на маркетплейсе Wildberries. Прежде всего, следует отметить, что все отзывы делятся на 5 уровней по степени негативности. Для обработки отзывов из личного кабинета Wildberries, были оставлены только отзывы уровня 1, 2 и 3, что говорит больше о негативности отзыва. Разметка данных, включая ее категорию, выполнялись вручную (см. рис. 2 и рис. 3).

2. Способ, использующий для формирования корпуса данных внешние источники с аналогичными данными. В этом случае датасет Nykaa-Cosmetics-Products-Review-2021 с отзывами на косметику (на английском языке) был скачан с сайта Kaggle.com. Полученные данные переведены на русский язык, представлены в виде таблицы, отсортированы по оценке, и вручную отнесены к определенной категории (см. рис. 4-6).

	A	B	E	G	
1	ID отзыва	Дата	Количество звезд	Текст отзыва	Имя
2	VKUIBYgBWg4f_fchGU8j	10/5/2023	5	Крем соответствует ожиданиям .пользуюсь три года .	
3	d5ykBIgBYE-HFOZ5xjFC	10/5/2023	5	Флакон пришёл в заводской упаковке. Дозатор рабочий, срок	Ника
4	moQxBIgBUYwa0QEfce8j	10/5/2023	5	Всегда актуальный продукт	Наталья
5	ey8jAogBG4PPbu1tpAp5	9/5/2023	5	Нормальная такая умывалка)	Оксана
6	HQi0AYgBZx8DR9RALIeN	9/5/2023	5	Хорошо упакован	
7	ybWPAYgBqVKcxP1mAPY6	9/5/2023	5	Доставка и сроки хорошие, брали в поездку, еще не пробовал	Вера
8	K_yOAYgBN75J-nvtwKRK	9/5/2023	5	Доставка и сроки хорошие, брали в поездку, еще не пробовал	Вера
9	7PBiAYgBCaW1RxrSe604	9/5/2023	5	Доставили быстро,чуть помятая коробка была,но это не столь	Сидика
10	EwhOAYgBZx8DR9RAT1uT	9/5/2023	3	Трудно оценить. Товар пришёл вскрытый. Кто ж знает, что туд	Ирина
11	K78sAYgBuWYO9ah2hyis	9/5/2023	5	Крем хорошый	Лилия
12	vun0AIgBlM6cTg_R_pSX	9/5/2023	5	Покупала по отзывам, хорошо упакован, пришёл быстро, Спас	Елена
13	zyvVAIgBNO444I63_wpm	9/5/2023	1	срок годности истек, ребенок ходил на пунк выдачи. теперь вс	Татьяна

Рисунок 2 – Данные из личного кабинета Wildberries

Текст отзыва	Категория
Не упакован, потек.	Упаковка
Обгорела в первый же день. Причем мазалась очень часто им	Эффект
Крем полное ..., перед этим был такой же есть с чем сравнивать, запах не тот, консистенция жирный.... Думала контрафакт, а не производство Россия, а предыдущий другой страны. Даже содрать не могут приличное сделать. Еле смыла, под тональный вообще не ложится, хотя предыдущий идеально заходил.... Не советую, разве только пятки мазать, что б не сгорели)	Полный набор
Ничего особенного,эффекта не поняла,больше брать не буду	Эффект
Возмущена. Крем свернулся. Не впитывается, даже спустя время кожа остаётся жирная,сверху белые жирные шарики. Негоден. Он в сроке по дате производства, но, видимо, хранился при высокой температуре. Выбросить только.	Консистенция
Со своей задачей не справился совсем. Наносился крем до выхода на пляж и обновлялся после каждого захода в море. В итоге нос, лоб и плечи сгорели. Кисти рук (где держится за коляску) у ребенка мазала бесконечно, тоже подгорели и шелушились.	Эффект

Рисунок 3 – Размеченные данные из личного кабинета Wildberries

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	product_id	"brand_name"	"review_id"	"review_title"	"review_text"	"author"	"review_date"	"review_rating"	"is_a_buyer"	"pro_user"	"review_label"	"prod			
2	781070	"Olay"	"16752142"	"Worth buying 50g one"	"Works as it claims. Could see the difference from the first day. Use it with Olay cleanser for best result										
3	781070	"Olay"	"14682550"	"Best cream to start ur day"	"It does what it claims . Best thing is it smoothens ur skin n makes it soft . I liked it"	"Amrit Neelam"									
4	781070	"Olay"	"15618995"	"perfect for summers dry for winters"	"I have been using this product for months now.. it is perfect for combination n oily skin a										
5	781070	"Olay"	"13474509"	"Not a moisturizer"	"i have an oily skin, while this whip acts as a great base or primer as it smoothens the skin but it does not m										
6	781070	"Olay"	"16338982"	"Average"	"It's not that good. Please refresh try for other products"	"Sukanya Sarkar"	"2020-12-22 15:24:35"	"2.0"	"True"	"False					
7	781070	"Olay"	"14549640"	"not good for oily skin"	"dz product z best for dry skin ...one of olay representative ...suggest me to buy dz..she said its all skin t										
8	739418	"Olay"	"16531371"	"All time favorite"	"This cream is just awesome, It makes my rough skin soft and smooth without leaving any oily effect an all! I										

Рисунок 4 – Данные с сайта Kaggle.com

B	C	D	E	F
brand_name	review_id	review_title	review_text	author
Olay	14549640	not good for oily skin	dz product z best for dry skin ...one of olay representative ...sug	Laxmi Basumatary
Olay	29118808	Horrible cream	Does nothing, only increases pimples	Gayatri Prakash
Olay	712286	Simply superbb	This is the first time am reviewing or commending a product. I n	sri lakshmi Rajesh
Olay	3207671	Worst product ever.	I have an acne prone skin. Thought of trying this. It gave me so	kritika sharma
Olay	4230636	Useless	Nothing works. It's advertisement is good and got tempted to u	Theirisa Mary
Olay	17519613	NOT EFFECTIVE	This product is not effective at all...its a waste of money.	faizah feroz
Olay	27938807	Very bad eye cream	My under eyes are dark and dry after using this cream..waste o	Koyel Bhunia
Olay	26462130	No change...	Wrost product.. cost is high.. waste of money	Nalini

Рисунок 5 – Обработанные данные с сайта Kaggle.com

	B	G	
1	review_te	Перевод текста	Категори
3	Does nothin	Ничего не делает, только увеличивает прыщи	Эффект
5	I have an ac	У меня кожа, склонная к прыщам. Думал о том, чтобы попробовать это. Это дало мне так много прорывов. Мое лицо выглядит, что у меня есть какая -то аллергия. Не подходит для жирной для комбинированной кожи.	Эффект
6	Nothing wor	Ничего не работает. Реклама хорошая, и у вас возникла искушение использовать эти продукты. Я использовал как дневной, так и ночной крем и никаких результатов. У меня есть проблема с пигментацией, и это не помогает. Единственная позитивная вещь, то, что текстура подходит моей коже. Пустая трата денег. Никогда не купит снова и не порекомендует другим купить.	Эффект
7	This product	Этот продукт совсем не эффективен ... это пустая трата денег.	Эффект
8	My under ey	Мои под глазами темные и сухие после использования этого крема ... Главный деньги ...	Эффект
		Ребята, пожалуйста, не покупайте этот продукт ... Я поделюсь, какой опыт у меня был после использования этого 2 -месячного, он просто высушивает мою область глаз, у меня никогда не было проблем с сухостью, потому что у меня есть комбинированная кожа, и кредит, что это заставило мою кожу склонной к морщин из -за сухости, и она никогда не смешивается должным	

Рисунок 6 – Размеченные данные с сайта Kaggle.com

3. Способ, использующий для формирования корпуса данных ChatGPT – языковую модель на базе искусственного интеллекта, разработанную OpenAI, с помощью которой можно генерировать творческие тексты (версия – бот в Telegram https://t.me/chatsgpts_bot). Для получения набора негативных отзывов (рекламаций) с различными категориями претензий, были сформированы запросы нескольких видов. Примеры запросов и ответов к ChatGPT показаны на рисунках 7 и 8.

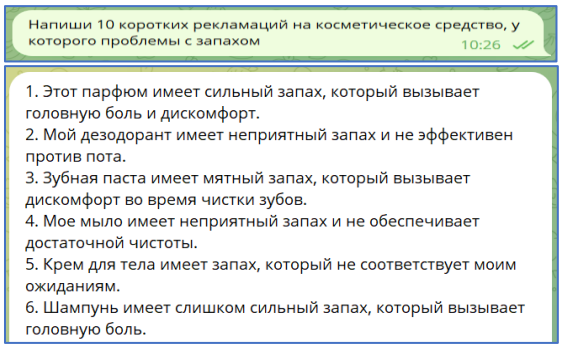


Рисунок 7 – Генерация первого вида отзывов с помощью нейросети

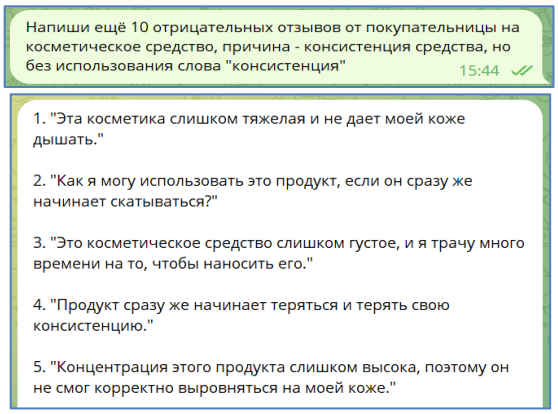


Рисунок 8 – Генерация второго вида отзывов с помощью нейросети

В качестве исходных данных получили набор в формате .csv из 1109 размеченных отзывов и рекламаций.

Анализ полученного корпуса данных

Одним из методов оценки качества данных является визуальный метод. Чтобы оценить сбалансированность созданного корпуса рекламаций, было рассчитано процентное соотношение по разным категориям текстов всего корпуса и группы с реальными рекламациями, поступившими в отдел продаж.

На рисунках 9 и 10 представлены зависимости количества встретившихся фактов по размеченным категориям для наборов из реальных данных и всего полученного корпуса.

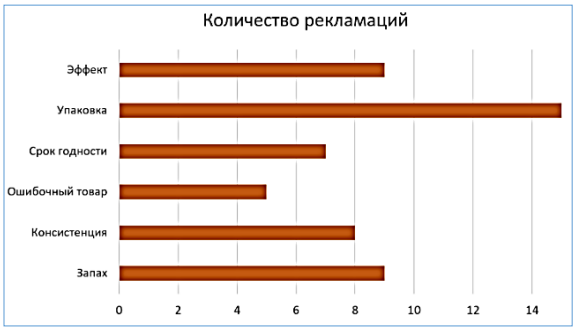


Рисунок 9 – Распределение рекламаций по категориям в наборе из реальных данных

Рисунок 10 – Распределение рекламаций по категориям в наборе из реальных данных

Результаты оценки сбалансированности корпуса представлены в таблице 1.

Таблица 1 – Процентное соотношение данных по категориям

Категория	Кол-во по корпусу	Процентное распределение по категориям	Кол-во по реал. данным	Процентное распределение по категориям	Погрешность
1-Запах	156	0,140667	9	0,169811	0,02914
2-Консистенция	159	0,143372	8	0,150943	0,00757
3-Ошибочный товар	139	0,125338	4	0,094347	0,04841
4-Срок годности	124	0,111812	7	0,132075	0,02026
5-Упаковка	342	0,308386	15	0,283019	0,02536
6-Эффект	189	0,170424	9	0,169811	0,00061

Отклонение представляет собой модуль разницы между процентными показателями текстов всего корпуса и текстов реальных рекламаций. Поскольку отклонения для всех типов категорий не превышают 5%, можно сделать вывод о сбалансированности корпуса и о возможности применения методов машинного обучения и автоматической обработки текстов к созданному корпусу.

Кодирование категориальных данных

В большинстве алгоритмов машинного обучения набор данных может содержать текстовые или категориальные значения (в основном не числовые значения). Несколько алгоритмов, таких как CatBoost [5], могут достаточно хорошо обрабатывать категориальные значения, но большинство алгоритмов работают лучше с числовыми данными. Следовательно, основная задача заключается в том, чтобы преобразовать текстовые категориальные данные в числовые. результат.

Есть два основных метода кодировки Label-Encoder и One-Hot Encoding, которые являются частью библиотеки Scikit-learn (Python) и используются для преобразования текстовых или категориальных данных в числовые данные.

Метод Label Encoder включает преобразование каждого значения в столбце в число. Категориальные признаки рекламаций сведены в таблицу 2.

Таблица 2 – Метод Label-Encoder

Категория (текст)	Категория (число)
запах	0
консистенция	1
ошибочный товар	2
срок годности	3
упаковка	4
эффект	5

При кодировании способом меток появляется проблема, связанная с использованием ряда чисел. Хотя нет никакой связи между различными категориями, создается впечатление, что категории ранжированы. Существует альтернативный метод, называемый «One-Hot-Encoding». В этом алгоритме каждое значение категории преобразуется в новый столбец, и столбцу присваивается значение 1 или 0 (см. табл. 3).

Хотя этот подход устраняет проблемы порядка, но приводит к значительному увеличению количества столбцов. Выбор между методами Label-Encoder и One-Hot-Encoding зависит от конкретного набора данных и модели, которую в дальнейшем планируется применить.

Таблица 3 – Метод One-Hot Encoding

Категория(текст)	0	1	2	3	4	5
0 - запах	1	0	0	0	0	0
1 - консистенция	0	1	0	0	0	0
2 - ошибочный товар	0	0	1	0	0	0
3- срок годности	0	0	0	1	0	0
4 - упаковка	0	0	0	0	1	0
5 - эффект	0	0	0	0	0	1

В случае, если категориальная особенность не является порядковой и количество категорий небольшое, для выбора правильной техники кодирования предпочтительным будет использование метода One-Hot Encoding.

Если же категориальная особенность является порядковой и количество категорий довольно велико метод One-Hot-Encoding может привести к высокому потреблению памяти и следует использовать метод Label-Encoder.

В рассматриваем случае использовался метод One-Hot Encoding (см. рисунок 11).

```
[ ]  nb_classes = 6
     y_test = utils.to_categorical(y_test_data, nb_classes)
     y_test

     array([[0., 1., 0., 0., 0., 0.],
            [0., 1., 0., 0., 0., 0.],
            [1., 0., 0., 0., 0., 0.],
            ...,
```

Рисунок 11 – Категории, записанные с помощью One Hot Encoding

Обработка датасета с помощью NLP

NLP или обработка естественного языка – это область искусственного интеллекта (ИИ), которая лежит на пересечении машинного обучения и математической лингвистики и помогает взаимодействовать между компьютерами и человеческим языком.

Прежде чем приступать к решению задач машинного обучения, в том числе и векторизации текстовых данных в NLP, имеющиеся данные корпуса следует предварительно обработать. При предварительной обработке данных их следует очистить и улучшить, чтобы модель могла работать более эффективно.

Основные методы предобработки текста изображены на рис. 12.

Для предобработки данных из рекламаций были использованы функции библиотек языка Python. В ходе выполнения лексического анализа документа сначала выполняется токенизация – процесс разбиения текста на текстовые единицы, такие как слова или предложения.

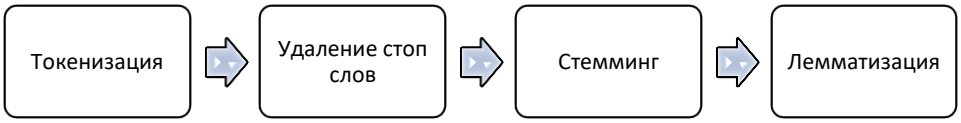


Рисунок 12 – Основные методы предобработки текстов процесса решения задач с МО и ИАД

Далее данные следует очистить от лишней информации, в том числе и от знаков пунктуации. В русском языке служебные части речи: предлоги, частицы и союзы служат только для связи слов в предложении. Кроме этого, есть слова, которые встречаются почти в каждом предложении, не несут большой информативной нагрузки и называются стоп-словами. Они являются шумом для последующего МО и плохо влияют на качество классификации.

Следующий этап – стемминг (stemming). Так как русский язык обладает богатой морфологической структурой, то для МО лучше привести их к одной форме для уменьшения размерности. В частности, он обрезает окончания слов. В Python-библиотеке NLTK для этого есть метод SnowballStemmer, который поддерживает русский язык. Проблемы могут возникнуть со словами, которые значительно изменяются в других формах. Поэтому лучше применить другой метод – лемматизацию [5].

На этапе лемматизации над словом можно провести морфологический анализ и выявить его начальную форму (это когда слова сводятся к их корневым формам для обработки). Для этого возможно воспользоваться Pymorphy2 – инструментом для морфологического анализа русского и украинского языков [6]. Метод Parse возвращает список объектов Parse, которые обозначают виды грамматических форм анализируемого слова.

Процесс NLP-предобработки данных

Алгоритм процесса предобработки представлен ниже.

1. Подключение Pymorphy2 – для морфологического анализа слов, pandas – для работы с данными и NLTK – для обработки текста.

```
!pip install pymorphy2

import nltk
import pandas as pd
import pymorphy2

from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

Рисунок 13 – Данные, считанные из файла

2. Считывание рекламации из файла. Результат изображен на рисунке 14.

```
claims = pd.read_csv('/content/DataForGoogleCol.csv', sep=';')
claims
```

	Текст	Категория
0	Данный продукт привел к появлению высыпаний на...	Эффект
1	Этот косметический продукт имеет неприятный за...	Запах
2	Я обнаружил грибок на своей помаде.	Эффект
3	Этот тональный крем меняется цветом на коже.	Эффект
4	У меня возник неприятный ощущения жжения и раз...	Эффект
...	...	...
1104	Не советую портить отдых из за этого масла	Эффект
1105	не работает распылитель	Упаковка
1106	не работает распылитель, возврат	Упаковка
1107	Не работает вообще, расход огромный, за два дн...	Эффект
1108	не работает средство	Эффект

1109 rows × 2 columns

Рисунок 14 – Данные, считанные из файла

3. Загрузка из NLTK модуля для стоп-слов.

Фрагмент кода загрузки модуля стоп-слов приведен на рисунке 15.

```
nltk.download('stopwords')

[nltk_data] Downloading package stopwords to /root/nltk_data.
[nltk_data]   Package stopwords is already up-to-date!
True
```

Рисунок 15 – Загрузка модуля стоп-слов

4. Сохранение в переменную одну из рекламаций и перевод её в нижний регистр. Фрагмент кода, удаляющий стоп-слова и знаки пунктуации приведен на рисунке 16.

```
[38] text_lower = text_sample.lower()
     text_lower

     'я очень разочаровалась от использования
      этого блеска для губ. и хотя цвет
      был точно таким, как я себе
```

Рисунок 16 – Преобразование регистра

5. Разбивка текста на токены (слова и символы) с помощью токенизатора Python-библиотеки NLTK. Результат токенизации приведен на рисунке 17.

```
[39] tokens = word_tokenize(text_lower)
     tokens

     ['я',
      'очень',
      'разочаровалась',
      'от',
      'использования',
      'этого',
      'блеска',
      'для',
      'губ',
      '.',
      'и',
      'хотя',
      'цвет',
      'был',
      'точно',
      'таким',
      ',',
      'как',
```

Рисунок 17 – Результат после этапа токенизации

6. Переход к этапу очистки данных, на котором происходит исключение стоп-слова из исходного текста. Библиотека NLTK имеет список стоп-слов на русском языке. Этот список можно скачать и использовать. Список же пунктуационных знаков, которые требуется удалять, необходимо создавать самим. Таким же образом можно расширять список стоп-слов. Для морфологического анализа используется метод из Pymorphy2. Фрагмент кода, удаляющий стоп-слова и знаки пунктуации приведен на рисунке 18.

```
[40] punctuation_marks = ['!', ',', '(', ')',
     ':', '-', '?', '.', '..', '...']
     stop_words = stopwords.words("russian")
     morph = pymorphy2.MorphAnalyzer()
```

Рисунок 18 – Результат после этапа очистки данных

7. Лемматизация данных. Результаты лемматизации некоторой произвольной фразы изображены на рисунке 19.

```
preprocessed_text = []

for token in tokens:
  if token not in punctuation_marks:
    lemma = morph.parse(token)[0].normal_form
    if lemma not in stop_words:
      preprocessed_text.append(lemma)

preprocessed_text

['очень',
 'разочароваться',
 'использование',
 'это',
 'блеск',
 'губа',
 'хотя',
 'цвет',
 'точно',
 'представлять',
 'запах',
 'напоминать',
 'химический',
 'лаборатория']
```

Рисунок 19 – Результат после этапа лемматизации

Выводы

В статье описаны основные этапы формирования и обработки корпуса данных с целью получения датасета для дальнейшего использования при решении задач машинного обучения.

При формировании корпуса данных решалась задача неполноты данных. При этом использовались различные варианты получения данных датасета: из рекламаций покупателей и отзывов пользователей фирмы; из отзывов об аналогичной продукции других производителей; генерация отрицательных отзывов с помощью нейронной сети. В результате был сформирован набор размеченных данных и выполнена оценка корпуса на сбалансированность.

С помощью NLP-методов предобработки подготовлен набор данных, который будет в

дальнейшем использован при векторизации данных и решении задачи классификации методами МО.

Литература

1. Глазкова, А.В. Формирование текстового корпуса для автоматического извлечения биографических фактов из русскоязычного текста // International Journal of Open Information Technologies. 2019. №1. URL: https://cyberleninka.ru/article/n/formirovanie-tekstovogo-korpusa-dlya-avtomaticheskogo-izvlecheniya-biograficheskih-faktov-iz-russkoyazychnogo-teksta (дата обращения: 10.04.2023).

2. Вовченко, В. О. Структурно-функциональная модель процесса анализа рекламаций / В. О. Вовченко, В. А. Светличная // Информатика, управляющие системы, математическое и компьютерное моделирование (ИУСМКМ-2022) : Материалы XIII Международной научно-технической конференции в рамках VIII Международного Научного форума Донецкой Народной Республики, Донецк, 25–26 мая 2022 года. – Донецк: Донецкий национальный технический университет, 2022. – С. 202-207.

3. Андриевская, Н. К. Онтологический подход в системах обработки данных научных и научно-образовательных организаций //Проблемы искусственного интеллекта. – 2020. – №. 1. – С. 23-36.

4. Датасеты для машинного обучения и анализа данных: что это, виды - где взять датасеты (yandex.ru). URL: https://practicum.yandex.ru/blog/dataset-dlya-mashinnogo-obucheniya-i-analiza/ (дата обращения: 10.04.2023).

5. What is One Hot Encoding? Why And When do you have to use it? [Электронный ресурс]/2019 г. — Режим доступа: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f

6. ML | Label Encoding of datasets in Python [Электронный ресурс]/2019 г. — Режим доступа: https://www. geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/

7. Главная страница Python-School. [Электронный ресурс] — URL: https://python-school.ru/nlp-vectorization-methods/ (Дата обращения: 18.03.2023).

Вовченко В. О., Светличная В. А., Андриевская Н. К. Формирование датасета для решения задач машинного обучения. *Статья посвящена описанию основных этапов формирования корпуса данных для машинного обучения, а также методов предобработки текстов. Приведены варианты решения таких проблем, как неполнота данных, очистка и преобразование данных. Выполнено кодирование категориальных данных. С помощью методов предобработки NLP подготовлен набор данных, который будет в дальнейшем использован при векторизации и решении задачи классификации методами машинного обучения.*

Ключевые слова: *датасет, машинное обучение, NLP, токенизация, лемматизация*

Vovchenko B. O., Svetlichnaya V. A., Andrievskaya N. K. Creating a Dataset for Machine Learning. *Formation of a Dataset for Machine Learning Problems. The article describes the main stages of data set formation for machine learning, as well as methods of text preprocessing. Variants of solving such problems as data incompleteness, data cleaning and transformation are given. Categorical data coding is performed. With the help of NLP preprocessing methods, a data set is prepared which will be further used in vectorization and solving the problem of classification by machine learning methods.*

Keywords: *dataset, machine learning, NLP, tokenization, lemmatization*

Статья поступила в редакцию 25.05.2023.
Рекомендована к публикации профессором Скобцовым Ю.А.

УДК 629.7.054.07

Особенности моделирования работы системы амортизации бесплатформенного инерциального измерительного прибора на языке Python и в среде Simulink

П.А. Илюшин, В.П. Наумченко, Д.Г. Пикунов, А.В. Соловьев
филиал АО «ЦЭНКИ» – «НИИ ПМ им. академика В.И.Кузнецова»
E-mail: P.Ilyushin@russian.space

Аннотация

В настоящей работе рассматривается процесс разработки системы амортизации и демпфирования прецизионного вибрационно-стойкого бесплатформенного инерциального измерительного прибора космического назначения. В рамках проведенных работ были созданы математические модели в MATLAB/Simulink и Python для решения линейной системы дифференциальных уравнений в символьном виде и нелинейной системы численными методами. Результаты частично подтверждены в рамках натурных испытаний.

Введение

В рамках разработки вибростойкого высокоточного инерциального измерительного прибора космического назначения (прибор) было обнаружено, что примененные в нем вибрационно-струнные акселерометры (ВСА) выходят из строя при максимальных режимах вибрационного воздействия, характерных для нештатного полета космического аппарата. Причем конструкция ВСА отработана и вносить в нее изменения недопустимо.

Возникла необходимость создания внутренней системы амортизации прибора с подвижным блоком чувствительных элементов (БЧЭ). В процессе предварительной проработки вопроса было обнаружено, что эффективность гашения колебаний в ВСА увеличивается с уменьшением собственной частоты амортизации БЧЭ. Однако, чем меньше частота амортизации, тем больше собственные перемещения БЧЭ и тем больше размер корпуса прибора и габариты необходимой системы демпфирования для уменьшения коэффициента усиления на резонансной частоте.

В итоге задача свелась к классической задаче поиска оптимального решения по двум критериям (перемещение ВСА и БЧЭ) при варьировании конструктивных параметров системы амортизации и демпфирования (САД), определяющих динамические характеристики системы.

Исследуемый динамический объект и его математическая модель

Ранее уже проводился цикл экспериментальных и теоретических исследований прибора-прототипа (прототип), в котором также применялись ВСА и БЧЭ с САД. В ВСА реализована собственная САД, предназначенная для защиты чувствительного элемента ВСА от внешних ударов. Однако, эта система при высокоэнергетических воздействиях может приводить к ударам подвижной части ВСА о его корпус. Для исключения этого в прототипе применен еще один амортизатор, представляющий площадку, связанную с корпусом прибора посредством равножестких пружин. Уменьшение колебаний самой площадки обеспечивается настроенными на частоту этих колебаний виброгасящими демпферами (ВД). В прототипе ВСА установлены соосно, в приборе ВСА расположены по конусу (**Ошибка! Источник ссылки не найден.**).

В ходе натурных и теоретических исследований было обнаружено, что исходная САД прототипа не позволяет обеспечить вибростойкость при требуемых воздействиях [1]. Было принято решение подобрать параметры САД, которые можно технически варьировать, исходя из заданных требований к перемещениям ВСА и БЧЭ.

Движение в САД прибора можно описать в виде системы из 8 уравнений в векторной форме (1):

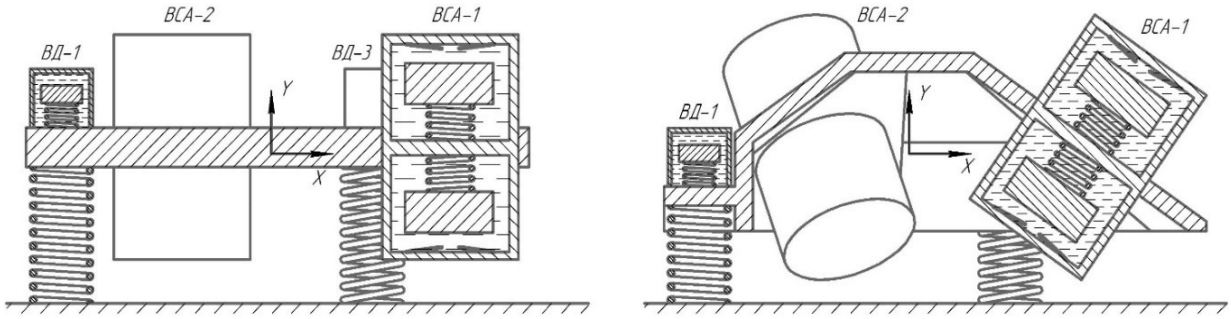


Рисунок 1 – Принципиальная схема САД (слева прототип, справа прибор)

$$\begin{cases} \boxed{\begin{array}{l} \mathrm{M_{bi}}\cdot\left(\ddot{\vec{\mathrm{R}}}_{\mathrm{bi}}-\vec{\mathrm{g}}\right)+\sum\limits_{\mathrm{i}=1}^{3}\left(\begin{array}{c}(\beta_{\mathrm{bi}}+\beta_{\mathrm{vi}})\cdot\left(\dot{\vec{\mathrm{R}}}_{\mathrm{bi}}-\dot{\vec{\mathrm{R}}}^{\mathrm{b}}_{\mathrm{in}}\right)+\mathrm{C_{bi}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{bi}}-\vec{\mathrm{R}}^{\mathrm{b}}_{\mathrm{in}}\right)-\\ -\mathrm{C_{di}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{di}}-\vec{\mathrm{R}}^{\mathrm{d}}_{\mathrm{bi}}\right)-\mathrm{H}_{\mathrm{ai}}^{-1}\cdot\mathrm{C_{ai}}\cdot\mathrm{H_{ai}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{ai}}-\vec{\mathrm{R}}^{\mathrm{a}}_{\mathrm{bi}}\right)+\vec{\mathrm{F}}_{\mathrm{fi}}\end{array}\right)=\vec{0}\\ \mathrm{M_{di}}\cdot\left(\ddot{\vec{\mathrm{R}}}_{\mathrm{di}}-\vec{\mathrm{g}}\right)+\beta_{\mathrm{di}}\cdot\left(\dot{\vec{\mathrm{R}}}_{\mathrm{d}}-\dot{\vec{\mathrm{R}}}^{\mathrm{d}}_{\mathrm{bi}}\right)+\mathrm{C_{di}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{di}}-\vec{\mathrm{R}}^{\mathrm{d}}_{\mathrm{bi}}\right)=\vec{0}\\ \mathrm{M_{ai}}\cdot\mathrm{H_{ai}}\cdot\left(\ddot{\vec{\mathrm{R}}}_{\mathrm{ai}}-\vec{\mathrm{g}}\right)+\beta_{\mathrm{ai}}\cdot\mathrm{H_{ai}}\cdot\left(\dot{\vec{\mathrm{R}}}_{\mathrm{ai}}-\dot{\vec{\mathrm{R}}}^{\mathrm{a}}_{\mathrm{bi}}\right)+\mathrm{C_{ai}}\cdot\mathrm{H_{ai}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{ai}}-\vec{\mathrm{R}}^{\mathrm{a}}_{\mathrm{bi}}\right)=\vec{0}\end{array}}\Bigg\rangle\,\mathrm{i}=1..3\\ \mathrm{J_{bi}}\cdot\ddot{\vec{\varepsilon}}_{\mathrm{b}}+\sum\limits_{\mathrm{i}=1}^{3}\left(\begin{array}{c}\vec{\mathrm{R}}_{\mathrm{bi}}\cdot\left[(\beta_{\mathrm{bi}}+\beta_{\mathrm{vi}})\cdot\left(\dot{\vec{\mathrm{R}}}_{\mathrm{bi}}-\dot{\vec{\mathrm{R}}}^{\mathrm{b}}_{\mathrm{in}}\right)+\mathrm{C_{bi}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{bi}}-\vec{\mathrm{R}}^{\mathrm{b}}_{\mathrm{in}}\right)\right]+\vec{\mathrm{r}}^{\mathrm{b}}_{\mathrm{fi}}\cdot\vec{\mathrm{F}}_{\mathrm{fi}}+\\ \vec{\mathrm{R}}_{\mathrm{ai}}\cdot\mathrm{H}_{\mathrm{ai}}^{-1}\cdot\mathrm{C_{ai}}\cdot\mathrm{H_{ai}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{ai}}-\vec{\mathrm{R}}^{\mathrm{a}}_{\mathrm{bi}}\right)+\vec{\mathrm{R}}_{\mathrm{di}}\cdot\mathrm{C_{di}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{di}}-\vec{\mathrm{R}}^{\mathrm{d}}_{\mathrm{bi}}\right)\end{array}\right)=\vec{0}\\ \left(\mathrm{J_{bi}}+\mathrm{M_{bi}}\cdot\vec{\mathrm{r}}^{\mathrm{b}\,2}_{\mathrm{fj}}\right)\cdot\ddot{\vec{\varepsilon}}_{\mathrm{b}}-\vec{\mathrm{R}}^{\mathrm{b}}_{\mathrm{fj}}\cdot\mathrm{M_{bi}}\cdot\left(\ddot{\vec{\mathrm{r}}}_{\mathrm{b}}-\vec{\mathrm{g}}\right)+\\ +\sum\limits_{\mathrm{i}=1}^{3}\left(\begin{array}{c}\left(\vec{\mathrm{R}}_{\mathrm{bi}}-\vec{\mathrm{R}}^{\mathrm{b}}_{\mathrm{fj}}\right)\cdot\left[(\beta_{\mathrm{bi}}+\beta_{\mathrm{vi}})\cdot\left(\dot{\vec{\mathrm{R}}}_{\mathrm{bi}}-\dot{\vec{\mathrm{R}}}^{\mathrm{b}}_{\mathrm{in}}\right)+\mathrm{C_{bi}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{bi}}-\vec{\mathrm{R}}^{\mathrm{b}}_{\mathrm{in}}\right)\right]+\\ +\left(\vec{\mathrm{R}}_{\mathrm{di}}-\vec{\mathrm{R}}^{\mathrm{b}}_{\mathrm{fj}}\right)\cdot\mathrm{C_{di}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{di}}-\vec{\mathrm{R}}^{\mathrm{d}}_{\mathrm{bi}}\right)+\left(\vec{\mathrm{R}}_{\mathrm{fi}}-\vec{\mathrm{R}}^{\mathrm{b}}_{\mathrm{fj}}\right)\cdot\vec{\mathrm{F}}_{\mathrm{fi}}+\\ +\left(\vec{\mathrm{R}}_{\mathrm{ai}}-\vec{\mathrm{R}}^{\mathrm{b}}_{\mathrm{fj}}\right)\cdot\mathrm{H}_{\mathrm{ai}}^{-1}\cdot\mathrm{C_{ai}}\cdot\mathrm{H_{ai}}\cdot\left(\vec{\mathrm{R}}_{\mathrm{ai}}-\vec{\mathrm{R}}^{\mathrm{a}}_{\mathrm{bi}}\right)\end{array}\right)=\vec{0}\Bigg\rangle\,\mathrm{j}=1..3\end{cases}\quad(1)$$

В формуле (1):

$\mathrm{in,a,b,d,v}$ – элементы конструкции: корпус, БЧЭ, ВСА, ВД и дополнительное вязкое трение, соответственно (индекс снизу – элемент, индекс сверху – относительно какого элемента считается перемещение);

$\mathrm{M},\beta,\mathrm{C},\mathrm{J}$ – масса, вязкость жесткость, момент инерции элемента;

$\ddot{\vec{\varepsilon}}_b,\dot{\vec{\varepsilon}}_b,\vec{\varepsilon}_b$ – угловые ускорение, скорость и поворот БЧЭ;

$\mathrm{H},\dot{\mathrm{H}}$ – матрица поворота элемента и ее производная с учетом $\dot{\vec{\varepsilon}}_{\mathrm{b}},\vec{\varepsilon}_{\mathrm{b}}$;

$\ddot{\vec{\mathrm{R}}},\dot{\vec{\mathrm{R}}},\vec{\mathrm{R}}$ – линейные ускорение, скорость и перемещение элемента с учетом H и $\dot{\mathrm{H}}$;

i – номер элемента; j – номер элемента силы трения; $\vec{\mathrm{F}}_{\mathrm{f}}$ – сила трения.

Первым этапом решения задачи стало рассмотрение упрощенной линейной динамической системы САД прототипа [1] в которой происходило движение центров масс БЧЭ, ВСА, ВД. Элементы считались идентичными, т.е. исследовалось 3 степени свободы (уравнения в (1), отмеченные рамкой, в скалярной форме). Затем была разработана нелинейная модель [3], в которой было учтено срабатывание упоров в ВСА при достижении заданного перемещения. Нелинейная модель также уточнена в части кинематики ВСА и БЧЭ. Перемещения стали трехмерными векторами, а скалярные параметры САД были преобразованы в формат матриц 3 на 3. В промежуточном варианте было добавлено демпфирование в виде силы сухого трения, приложенной к центру масс БЧЭ и не вызывающей угловых колебаний. В окончательной модели три силы трения приложены в трех различных точках БЧЭ.

Средства и методы моделирования

Линейная версия модели была разработана в Python, решение производилось символьным методом посредством библиотеки SymPy. К системе уравнений было применено преобразование Лапласа, перемещения БЧЭ и ВСА были представлены в виде передаточных

функций. Таким образом оценивался максимальный коэффициент усиления при варьировании параметров САД и принималось решение о соответствии комплекса параметров заданным требованиям.

Для удобства анализа модели реализован графический интерфейс, регуляторы позволяют настраивать режим, варьируемые параметры и свойства рассматриваемой системы. Для автоматического поиска решений реализован метод перебора и генетический алгоритм.

Графики с результатами отображаются в окне, реализовано сохранение и построение по сохраненным рисункам gif-файла (Рисунок 1).

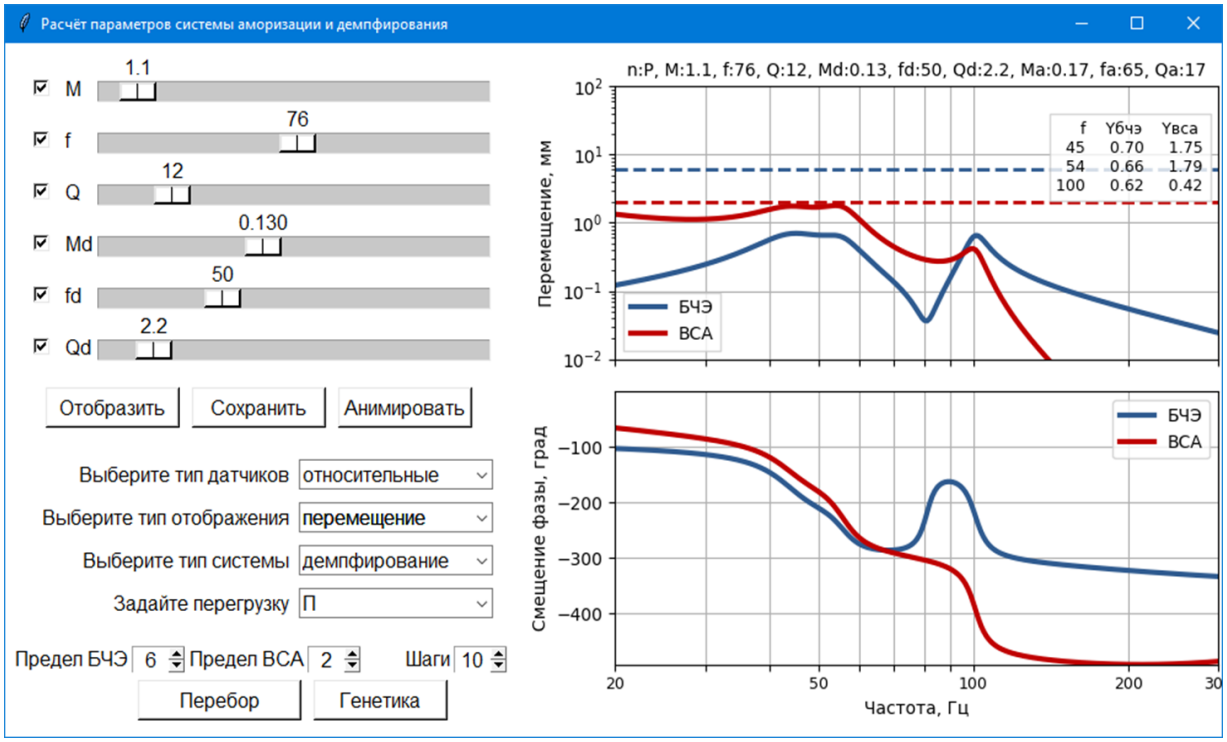


Рисунок 1 – Графический интерфейс управления вычислениями для линейной модели

Промежуточная нелинейная модель была разработана в MATLAB/Simulink, решение искалось встроенными численными методами. Модель собиралась из базовых блоков. В Simulink производилось моделирование, дискретизации и вывод результатов. В MATLAB осуществлялся запуск модели, перебор параметров и сохранение графиков. Было добавлено сохранение файла с итоговым результатом моделирования по всем полученным в запуске замерам.

Окончательная нелинейная модель была реализована в Python, решение искалось при помощи модуля Solve библиотеки SciPy. Система была предварительно преобразована в формат Коши. В этой модели были получены преимущества, характерные для скрипта с линейной моделью в Python: гибкость, производительность, удобство доработки и интегрирования в другие подсистемы. Сейчас управление происходит в формате корректировки скрипта в IDLE перед запуском, однако планируется интегрировать подсистему решения системы уравнений в разработанный ранее визуальный интерфейс.

В нелинейной модели на Python реализовано отображение тех же самых результатов, что и в модели MATLAB. Добавлены графики угловых колебаний, а на графиках АЧХ приводятся еще и характеристики пиков (**Ошибка! Источник ссылки не найден.**). Файл с результатами не перезаписывается, а автоматически дополняется при новых запусках моделирования.

Оценка методов моделирования

Нелинейная модель в MATLAB/Simulink обладает преимуществом в том, что писать код для решения системы уравнений не нужно. При этом программировать перебор параметров и сохранение результатов все-равно приходится. Есть и некоторые ограничения с запуском исполняемого скрипта MATLAB на разных ПК. В остальном нелинейная модель в MATLAB/Simulink уступает модели в Python, позволяющей более гибко настраивать отображение результатов и сам процесс вычислений. Доступны библиотеки для применения недоступных в MATLAB методов и

проведения реализованных в нем возможных вычислений.

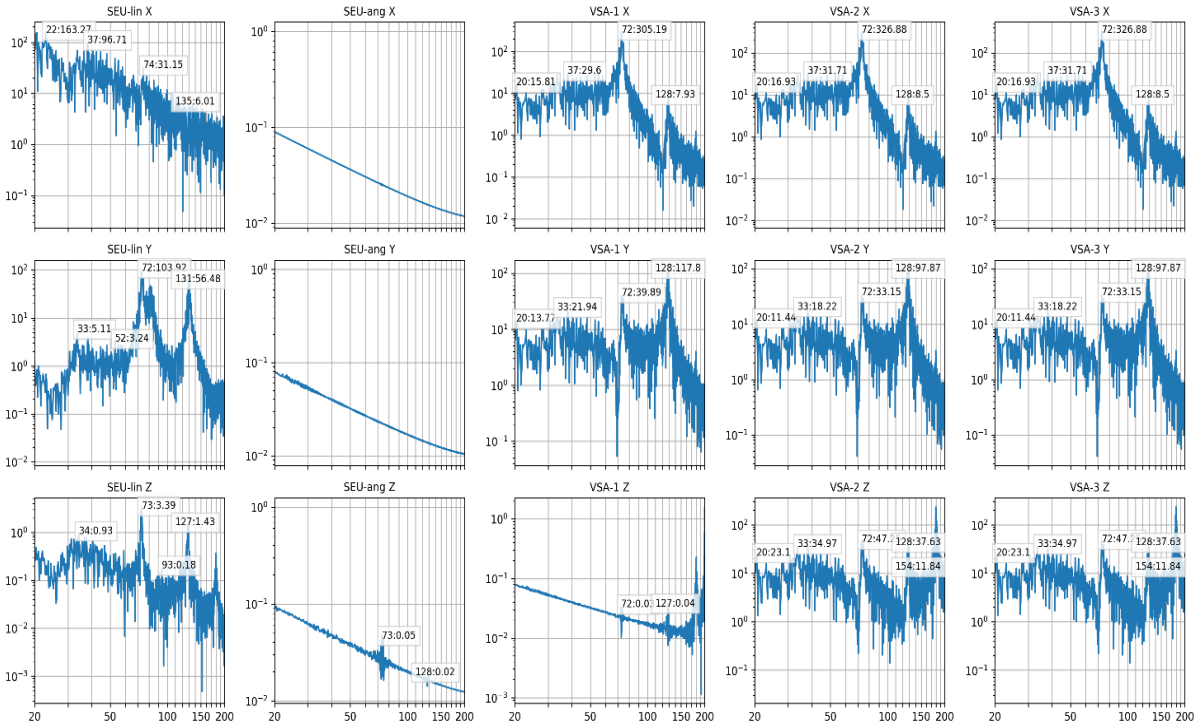


Рисунок 3 – Сохраняемый результат по результатам вычисления нелинейной модели

Также вычисления в MATLAB/Simulink шли быстрее, вероятно это связано с использованием метода RK45 [4].

В Python использовался метод LSODA, полезный для жестких задач, в отличие от метода RK45 в MATLAB/Simulink, предназначенного для решения нежестких задач.

Движение САД является жесткой задачей, для которой характерны множественные динамические процессы в разных временных масштабах.

Применяемый метод LSODA может автоматически переключаться между методом предсказания-коррекции Адамса (для нежестких задач) и методом BDF (для жестких задач).

Таким образом выбранный метод, при прочих равных точнее, надежнее, но потенциально медленнее нежестких методов. При замене метода на жесткий BDF в MATLAB/Simulink замер происходит заметно дольше, чем в Python.

Выводы

Подводя итог в части работы по подбору характеристик САД можно отметить, что при усложнении системы диапазон допустимых значений параметров сужался. Но в любом случае на текущий момент получены группы решений, обеспечивающие заданные требования по перемещениям в системе. Результаты частично подтверждены в рамках натурных испытаний.

Литература

1. Илюшин, П. А. Анализ влияния параметров конструкции инерциального прибора на его динамические характеристики при внешнем вибрационном возмущении / П. А. Илюшин, В. П. Наумченко // XLVI Академические чтения по космонавтике: Сборник тезисов, посвященные памяти академика С.П. Королёва и других выдающихся отечественных ученых-пионеров освоения космического пространства. В 4-х томах, Москва, 25–28 января 2022 года. Том 3. – Москва: Издательство МГТУ им. Н. Э. Баумана, 2022.

2. Илюшин, П. А. Моделирование работы линейной системы амортизации и демпфирования бесплатформенного инерциального измерительного прибора / П. А. Илюшин, В. П. Наумченко, С. А. Максимов [и др.] // Авиация и космонавтика: тезисы 21ой международной конференции, Москва, 21–25 ноября 2022 года / Московский авиационный институт (национальный исследовательский университет). – Москва: Издательство "Перо", 2022.

3. Илюшин, П. А. Исследование обеспечения стойкости к внешним вибрационным возмущениям бесплатформенного инерциаль-

ного измерительного прибора при помощи нелинейных элементов системы амортизации / П. А. Илюшин, В. П. Наумченко, Д. Г. Пикунов, А. В. Соловьев // Молодежь. Техника. Космос: труды четырнадцатой общероссийской молодежной научно-технической конференции: в 4 т., Санкт-Петербург, 23–27 мая 2022 года.

4. Медведева, Н. В. Сравнение численных методов решения задачи Коши для обыкновенных дифференциальных уравнений / Н. В. Медведева, Е. С. Скряга Е.С. // Международный студенческий научный вестник, 2018. – № 2. - URL: https://eduherald.ru/ru/article/ view?id= 18343 (дата обращения: 12.04.2023).

Илюшин П.А., Наумченко В.П., Пикунов Д.Г., Соловьев А.В. Особенности моделирования работы системы амортизации бесплатформенного инерциального измерительного прибора на языке Python и в среде Simulink. *В настоящей работе рассматривается процесс разработки системы амортизации и демпфирования прецизионного вибрационно-стойкого бесплатформенного инерциального измерительного прибора космического назначения. В рамках проведенных работ были созданы математические модели в MATLAB/Simulink и Python для решения линейной системы дифференциальных уравнений в символьном виде и нелинейной системы численными методами. Результаты частично подтверждены в рамках натурных испытаний.*

Ключевые слова: *Амортизация, демпфирование, БИБ, моделирование, Python*

Ilyushin P.A., Naumchenko V.P., Pikunov D.G., Solovyov A.V. Features of modeling the operation of the shock absorption system of a free-form inertial measuring device in Python and in the Simulink environment*. In this paper, we consider the design of shock absorption and damping system for precious space strapdown inertial measurement unit for spacecraft. As part of the work, we create mathematical models in MATLAB/Simulink and Python to solve the linear system of differential equations in symbolic form and nonlinear system by numerical methods. The results were partially confirmed in the framework of field tests.*

Keywords: *Shock-absorption, dampening, SIMU, simulation, Python*

Статья поступила в редакцию 04.05.2023
Рекомендована к публикации профессором Павлышом В. Н.

УДК 669.531.12

Моделирование процесса окислительного обжига цинкового концентрата в среде Python 3.0

Р. В. Куртенков, В. А. Слободин, Е. В. Сизякова
Санкт-Петербургский горный университет
Кафедра металлургии
E-mail: victorslobodin2002@mail.ru

Аннотация
В работе рассматривается технология первого металлургического передела цинкового концентрата, а именно окислительного обжига. Разработана программа для моделирования процесса обжига цинкового концентрата, описывающая движение материальных потоков. Помимо расчётных массовых и процентных значений она позволяет получить графическую информацию о составе продуктов. В дальнейшем планируется написание подобных программ, описывающих движение материальных потоков, тепловых балансов, кинетики, термодинамики других металлургических процессов.

Введение

К настоящему времени человечество освоило производство более 70 металлов. Технологическая цепочка получения некоторых из них может включать десятки процессов, проводимых в разнообразных металлургических аппаратах. Технологическим процессом (ТЕП) называется совокупность всех процессов, реализуемых в аппарате при переработке исходного сырья в конечные продукты [1].

Исследование сложных объектов с помощью их упрощённых моделей является очень плодотворным и широко используется в различных отраслях знаний.

Модель – это объект, который отражает основные, наиболее характерные черты изучаемого предмета или процесса, интересующие исследователя в данный момент времени. Она должна отражать не все свойства объекта, а только необходимые для решения конкретной задачи. Следовательно, в зависимости от целей исследования, для одного и того же объекта могут быть созданы различные модели [1].

Актуальность работы

Процесс окислительного обжига цинкового концентрата применяется при производстве цинка как по традиционной пирометаллургической, так и по развивающийся в современном мире гидрометаллургической технологи [2,3], а расчёты, необходимые для проведения данной операции очень трудоёмки и занимают большое количество времени, например, студенты, изучающие дисциплину «Металлургическая теплотехника и основы печных технологий» на расчёт материального баланса процесса обжига тратят от 8 до 10 академических часов. Исследование окислительного обжига является необходимым для получения профессиональных навыков студентами, изучающих технологии металлургии цветных металлов и пирометаллургическое оборудование. Для улучшения качества образовательного процесса при изучении данной темы кафедрой металлургии было предложено разработать программное приложение в среде Python 3.0 для расчета материальных потоков процесса обжига цинкового концентрата.

Цели, задачи и методы исследования

Целью исследования является компьютерное моделирование процесса окислительного обжига цинкового концентрата. Исходя из поставленной цели, были сформулированы следующие задачи:

1. Разработка и отладка программного кода для расчёта рационального состава исходного концентрата, рациональных составов продуктов обжига: огарка и пыли, состава отходящих газов и материального баланса процесса.
2. Разработка и отладка программного кода для построения круговых диаграмм элементных и вещественных составов концентрата, огарка и пыли и состава отходящих газов.

Они были достигнуты посредством следующих методов: статистическая обработка данных, сравнительный анализ, идеализация,

аналогия, обобщения и математического моделирования. Использованы теоретические зависимости по процессу окислительного обжига, описанные в Диомидовский Д.А. в работе «Расчеты пиропроцессов и печей цветной металлурги» [4]. Моделирование проходило в среде Python 3.0 по методикам, описанным Шариков Ю.В. «Моделирование процессов и объектов в металлургии» [1]. Для ввода и вывода численных значений использованы книги Excel. Для решения поставленных задач был разработан следующий укрупнённый алгоритм (рис. 1).

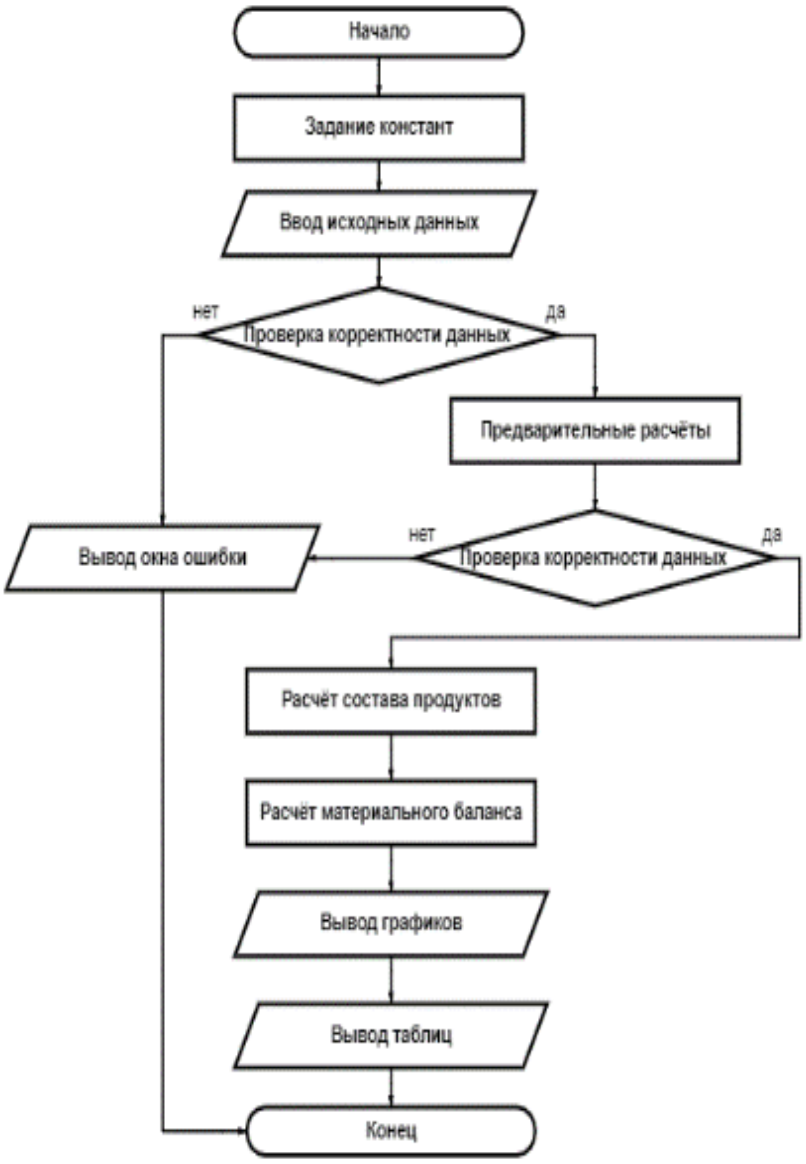


Рисунок 1 –Алгоритм работы программы

Теория процесса окислительного обжига цинкового концентрата

Цинк – светло-серый металл с синеватым оттенком. На воздухе он покрывается пленкой основного карбоната $ZnCO_3 \cdot 3Zn(OH)_2$ серого цвета, которая является весьма плотной и хорошо защищает цинк от дальнейшей коррозии [2].

Цинк широко используется для защиты железа от коррозии (30-60 % в разных странах от общего потребления). Наиболее распространенными сплавами, содержащими цинк, являются латуни и сплавы для литья под давлением. Оксид и сульфид цинка используются в качестве пигментов.

Существуют два типа цинксодержащих руд: сульфидные и оксидные. Главными природными сульфидными минералами цинка являются сфалерит ZnS и марматит (Zn, Fe)S. Для цинка характерна связь в рудах со свинцом, часто и с медью [2].

Средний химический состав сульфидных цинковых концентратов, являющихся основным природным сырьем для производства цинка, %: Zn 45-60, Pb 0,1-3,0, Cu 0,2-3,0, Cd 0,1-0,5, Fe 5-13, S 29-35, SiO_2 0,4-4. Их можно перерабатывать как пирометаллургическим, так и гидрометаллургическим методом. В настоящее время более 80 % от общего производства цинка приходится на гидрометаллургическую технологию [3]. Окислительный отжиг – первая операция в обоих методах переработки цинковых концентратов, целью которой является перевод сульфида цинка и сульфидов других металлов в форму оксидов.

При окислении сульфидов металлов выделяется большое количество тепла, что обеспечивает возможность проведения процесса обжига без других источников энергии [2].

Результаты

Разработанная программа использует модель с заданным минералогическим составом, включающим в себя:

- сфалерит (ZnS),
- галенит (PbS),
- халькопирит ($CuFeS_2$),
- гриконит (CdS),
- пирит (FeS_2),
- пирротин (Fe_7S_8),
- кварц (SiO_2),
- корунд (Al_2O_3) [4].

Программа использует следующие исходные данные:

- элементный состав сухого концентрата;
- производительность печи по сухому концентрату;
- влажность концентрата;
- коэффициент избытка воздуха;
- содержание кислорода в воздухе

Разработанная программа обеспечивает получение следующих выходных данных:

- таблица рационального состава исходного концентрата;
- таблица материального баланса (рис. 2);
- таблица рационального состава огарка (рис. 3);
- таблица рационального состава пыли;
- таблица состава отходящих газов;
- значение степени десульфуризации;
- значение удельного количества воздуха, требуемое для проведения процесса.

А также перечень графического материала:

- круговая диаграмма вещественного состава исходного концентрата;
- круговая диаграмма элементного состава исходного концентрата;
- круговая диаграмма вещественного состава огарка (рис. 4);
- круговая диаграмма элементного состава огарка (рис.4);
- круговая диаграмма вещественного состава пыли;
- круговая диаграмма элементного состава пыли;
- круговая диаграмма состава отходящих газов.

Материальный баланс

№ п/п	Приход	кг	%
1.	***Концентрат:***	191,489	31,037
	Zn	90	47
	Pb	2,7	1,41
	Cu	1,8	0,94
	Cd	0,54	0,282
	Fe	14,4	7,52
	S	57,6	30,08
	H2O	11,489	6
	SiO2	5,04	2,632
	Al2O3	5,4	2,82
	прочие	2,52	1,316
2.	***Воздух:***	425,473	68,963
	O2	97,859	23
	N2	327,614	77
	Итого:	**616,962**	**100**

№ п/п	Расход	кг	%
1.	***Огарок***	92,713	15,062
	Zn	54	57,954
	Pb	1,62	1,739
	Cu	1,08	1,159
	Cd	0,324	0,348
	Fe	8,64	9,273
	S(S)	0,371	0,398
	S(SO4)	0,927	0,995
	SiO2	3,024	3,245
	Al2O3	3,24	3,477
	прочие	1,512	1,623
2.	***Пыль***	64,473	10,474
	Zn	36	55,01
	Pb	1,08	1,65
	Cu	0,72	1,1
	Cd	0,216	0,33
	Fe	8,64	8,802
	S(S)	0,322	0,493
	S(SO4)	1,934	2,956
	SiO2	2,016	3,081
	Al2O3	2,16	3,301
	прочие	1,008	1,54
3.	***Отходящие газы***	458,342	74,463
	SO2	107,981	23,559
	O2	97,859	2,456
	N2	327,614	71,478
	H2O	11,489	2,507
	Итого:	**615,528**	**100**

Рисунок 2 - Выходные данные программы. Таблица результатов расчёта материального баланса

Рациональный состав огарка

Соединения	Zn		Pb		Cu		Cd		Fe		S(SO4)		S(S)		O2		SiO2		Al2O3		Прочие		Итого	
	кг	%	кг	%	кг	%	кг	%	кг	%	кг	%	кг	%	кг	%	кг	%	кг	%	кг	%	кг	%
ZnO	51,704	55,489													12,655	13,581							64,358	69,07
ZnSO4	1,54	1,653									0,756	0,811			1,508	1,618							3,804	4,083
ZnS	0,756	0,811											0,371	0,398									1,127	1,209
PbO			0,81	0,869											0,063	0,067							0,873	0,936
PbSO4			0,81	0,869							0,125	0,135			0,25	0,269							1,186	1,272
CdO							0,162	0,174							0,023	0,025							0,185	0,199
CdSO4							0,162	0,174			0,046	0,05			0,092	0,099							0,3	0,322
Cu2O					1,08	1,159									0,136	0,146							1,216	1,305
Fe2O3									8,64	9,273					3,713	3,985							12,353	13,257
SiO2																	3,024	3,245					3,024	3,245
Al2O3																			3,24	3,477			3,24	3,477
Прочие																					1,512	1,623	1,512	1,623
Итого	**54**	**57,954**	**1,62**	**1,739**	**1,08**	**1,159**	**0,324**	**0,348**	**8,64**	**9,273**	**0,927**	**0,995**	**0,371**	**0,398**	**18,439**	**19,79**	**3,024**	**3,245**	**3,24**	**3,477**	**1,512**	**1,623**	**93,177**	**100**

Рисунок 3 Выходные данные программы. Таблица результатов расчёта рационального состава огарка.

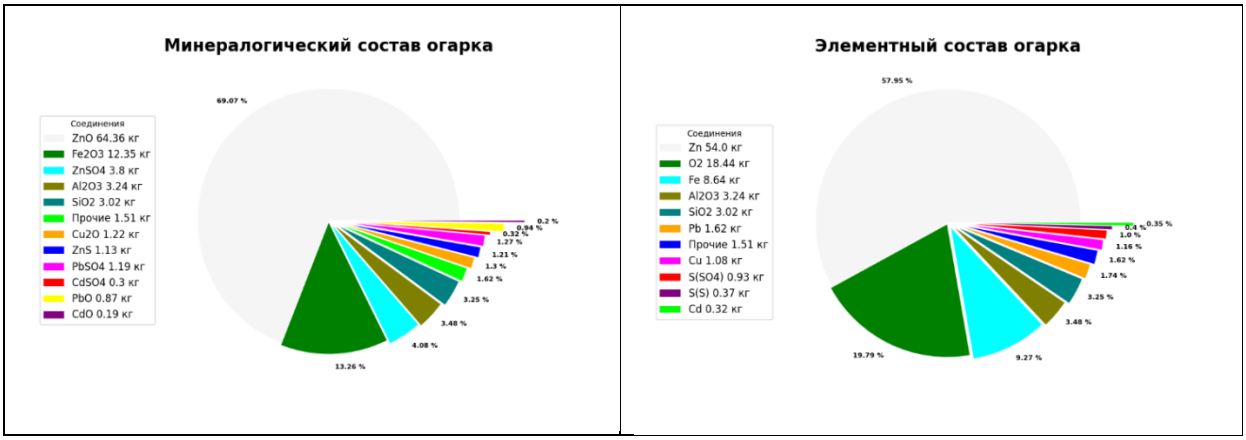


Рисунок 4 - Выходные данные программы. Круговые диаграммы минералогического и элементного составов цинкового огарка

Заключение

Приложение получило положительную оценку преподавателей кафедры металлургии Санкт-Петербургского горного университета. Применение программы планируется для обучения студентов бакалавров по специальности 15.03.04 - Автоматизации технологических процессов и производств, профиль: «Автоматизация технологических процессов и производств в металлургической промышленности» в рамках дисциплины «Пирометаллургическое оборудование» и по специальности 22.03.02 - Металлургия, профиль: «Металлургия цветных металлов» в рамках дисциплины «Металлургическая теплотехника и основы печных технологий». Студенты на практических занятиях получают задание в виде расчета теплового баланса процесса, печи для проведения обжига и системы газоочистки.

Для выполнения указанных расчётов им необходимо предварительно рассчитывать материальный баланс, количество и состав газовой фазы, хотя внимание на этот аспект обращено на других дисциплинах.

С помощью разработанной программы студенты смогут индивидуально рассчитывать данные по материальным потокам процесса и оперативно переходить непосредственно к расчету по теме практического занятия. Также возможно применение программы для разработки вариантов расчётно-графической работы (курсовой работы) по предмету «Теория пирометаллургических процессов». В дальнейшем планируется написание подобных программ, описывающих движение материальных потоков, тепловых балансов, кинетики, термодинамики других металлургических процессов.

Апробация работы

По итогам работы было получено свидетельство на программу для ЭВМ №2023612749 «Программа для моделирования процесса обжига цинкового концентрата», авторы: Слободин В.А., Куртенков Р.В., Сизякова Е.В. [5].

Литература

1. Шариков, Ю. В. Моделирование процессов и объектов в металлургии: учеб. пособие / Ю. В. Шариков, И. Н. Белоглазов, А.Ю. Фирсов. - Санкт-Петербургский государственный институт (технический университет). СПб, 2006. – 83 с.

2. Орлов, А. К. Основы производства и обработки металлов: Учебное пособие / А. К. Орлов, Г. В. Коновалов. - Санкт-

Петербургский государственный горный институт (технический университет). СПб, 2006. - 115 с.

3. Shao, S., Ma, B., Wang, C. et al. A Review on the Removal of Magnesium and Fluoride in Zinc Hydrometallurgy. J. Sustain. Metall. 8, 25–36 (2022). https://doi.org/10.1007/s40831-022-00500-4

4. Диомидовский, Д. А. Расчеты пиропроцессов и печей цветной металлурги: [Учеб. пособие для металлургич. вузов и фак.] / Д. А. Диомидовский, Л. М. Шалыгин, А. А. Гальнбек, И. А. Южанинов ; Под науч. ред. проф. д-ра техн. наук Д. А. Диомидовского. - Москва: Металлургиздат, 1963. - 459 с.

5. Свидетельство о государственной регистрации программы для ЭВМ № 2023612749 Российская федерация. Программа для моделирования процесса обжига цинкового концентрата : заявлено 20.01.2023 : опубликовано 07.02.2023 / Слободин В.А., Куртенков Р.В., Сизякова Е.В. ; правообладатель федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский горный университет», Бюл. № 2. – Зарегистрировано в Реестре программ для ЭВМ.

Куртенков Р.В., Слободин В.А., Сизякова Е.В. Моделирование процесса окислительного обжига цинкового концентрата в среде Python 3.0. *В работе рассматривается технология первого металлургического передела цинкового концентрата, а именно окислительного обжига. Разработана программа для моделирования процесса обжига цинкового концентрата, описывающая движение материальных потоков. Помимо расчётных массовых и процентных значений она позволяет получить графическую информацию о составе продуктов. В дальнейшем планируется написание подобных программ, описывающих движение материальных потоков, тепловых балансов, кинетики, термодинамики других металлургических процессов*

Ключевые слова: *технология, цинковый концентрат, окислительный обжиг, программа моделирования*

Kurenkov R.V., Slobodin V.A., Sizyakova E.V. Modeling of the process of oxidative firing of zinc concentrate in Python 3.0 environment. *The paper considers the technology of the first metallurgical conversion of zinc concentrate, namely oxidative firing. A program has been developed for modeling the process of firing zinc concentrate, describing the movement of material flows. In addition to the calculated mass and percentage values, it allows you to get graphical information about the composition of products. In the future, it is planned to write similar programs describing the movement of material flows, thermal balances, kinetics, thermodynamics of other metallurgical processes.*

Keywords: *technology, zinc concentrate, oxidative roasting, simulation program.*

Статья поступила в редакцию 07.05.2023
Рекомендована к публикации профессором Павлышом В. Н.

УДК 62-5, 681.5.015, 004.942

Применение методов анализа данных для определения наиболее популярных функций приложения по журналу действий пользователя

А.А. Личман, О.Ю. Чередникова
Донецкий национальный технический университет
E-mail: anton.lichman@yandex.ru

Аннотация
Рассмотрены методы решения задачи определения наиболее часто используемых функций приложения. Предложено использовать для этой цели нейронные сети. Выполнен анализ существующих нейронных сетей и особенностей их применения для различных целей. Предложен и реализован метод определения наиболее часто используемых функций приложения на основе нейронной сети Кохонена и прикладного пакета Deductor для предобработки данных. Это позволит разработчикам программных продуктов модернизировать уже готовые версии под новые потребности.

Введение

Одним из начальных этапов разработки или модернизации программного обеспечения (ПО) должно быть определение его функциональных потребностей. Для этого обычно анализируют текущую версию ПО или работу приложений, выполняющих похожие задачи. Для производственных целей часто существует многофункциональное ПО, которое выполняет сложные расчеты и приобретается на платной основе. Однако для многих конкретных задач предприятия достаточно небольшого набора функций. Например, часто используемое в геологии приложение «Micromine 2014», которое выполняет сложные расчеты, аналитику, часто применяют только в качестве визуализатора объемных моделей. В этом случае возможно рациональнее разработать собственное приложение, которое будет работать быстрее за счет меньших функциональных возможностей.

Для приложений, которые поддерживают ведение журнала действий пользователя, узнать требуемый пользователю функционал возможно, выполнив анализ журнала.

В настоящий момент актуальной является задача автоматизации анализа данных.

Исследование существующих решений анализа данных

Многие ведущие кампании заинтересованы в автоматизации анализа данных, в частности для выполнения анализа данных, получаемых от пользователей. Если объем данных небольшой, его возможно проанализировать вручную или программно с помощью офисных прикладных пакетов. Однако такие корпорации как Valve Steam и прочие пользуются нейронными сетями для анализа данных. Преимуществом нейронных сетей является скорость и качество их работы, а недостатком сложность реализации.

Еще с 2017 года Microsoft стали лидерами в Data mining за счет использования нейросетей для определения наиболее часто используемых пользователем функций и разработки на основе анализа новых версий операционной системы.

Алгоритмы с применением нейросетей, такие как метод ограниченного перебора приносят своим создателям коммерческую выгоду.

Компания WizSoft, например, разработала систему анализа данных WizWhy, основанную на алгоритме ограниченного перебора нейронных сетей, усовершенствовав этот метод при помощи алгоритма «Априори». Достоинством системы является простота в использовании и минимизация субъективных причин. Однако, главный ее недостаток - неспособность находить логические правила, содержащие более 6 элементарных событий [1].

Поэтому разработка методов анализа данных в настоящий момент остается актуальной и востребованной.

Постановка задачи

Технологии анализа данных (Data mining) применяют в различных отраслях человеческой деятельности

Целью применения Data Mining при анализе действий пользователя ПО является обнаружение наиболее частых действий, чтобы оптимально определить необходимый функционал разрабатываемого или модернизируемого ПО.

В работе предлагается решение следующих научных задач:

- исследование методов анализа данных;
- анализ типов нейросетей;

- реализация анализа журнала действий пользователя для определения наиболее частых действий на основе нейросети Кохонена.

Общие методы анализа данных

Помимо метода нейронных сетей следует выделить следующие методы анализа данных: деревья решений, генетические алгоритмы, нечеткая логика, алгоритмы ограниченного перебора, эволюционное программирование, системы рассуждения на основе аналогичных случаев, индукция правил, анализ с избирательным действием, логическая регрессия, алгоритмы определения ассоциаций и последовательностей, визуализация данных, комбинированные методы [2, 3, 6].

В технологии анализа данных (Data mining) большинство методов известны. Научной новизной является их адаптация под реализацию конкретных задач, благодаря развитию технологий последних лет.

Основная часть методов data mining была разработана в рамках теории искусственного интеллекта.

Метод нейронных сетей обычно используется для классификации, кластеризации, прогнозирования и распознавания образов. Для решения рассматриваемой в статье задачи нейросеть выполняет классификацию задач производства, основанную на анализе действий оператора, а также прогнозирования, чтобы делать приложение с заделом на возможные будущие задачи, или задатки под их выполнение [4].

Модель нейронной сети может быть следующих типов:

1) сети прямого распространения (backpropagation): одна из наиболее распространенных архитектур, в основном используется в таких областях, как прогнозирование и распознавание образов;

2) сети с обратной связью: такие, как дискретная модель Хопфилда, в основном используется для оптимизации вычислений и ассоциативной памяти;

3) самоорганизующиеся сети: включают модели адаптивной резонансной теории (ART) и модели Кохонена, в основном используется для кластерного анализа.

В настоящее время при анализе в data mining используются нейронные сети прямого распространения. Их недостатком является медленный темп обучения, а также высокий риск попасть в локальный минимум, из-за чего параметры обучения определить трудно.

Ввиду этих проблем многие перешли к методу объединения искусственных нейронных сетей с генетическими алгоритмами и достигли лучших результатов.

Одно из главных преимуществ нейронных сетей состоит в том, что они могут аппроксимировать любую непрерывную функцию, что позволяет исследователю не принимать заранее какие-либо гипотезы относительно модели. К существенным недостаткам нейронных сетей можно отнести тот факт, что окончательное решение зависит от начальных установок сети и его практически невозможно интерпретировать в традиционных аналитических терминах, что в целом не сильно мешает.

Процесс анализа данных, основанный на нейронной сети

Процесс анализа данных (data mining) может быть представлен тремя основными фазами:

- подготовка данных;
- анализ данных;
- выражение и интерпретация результатов (рис. 1).

Интеллектуальный анализ данных, основанный на нейронной сети, состоит из: подготовки данных, извлечения правил и оценки правил, то есть также трех этапов, как показано на рис. 2.

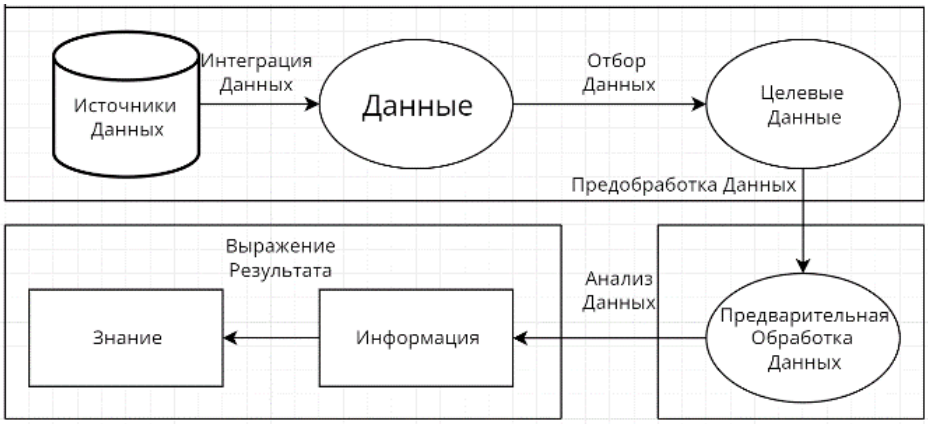


Рисунок 1 – Фазы процесса анализа данных

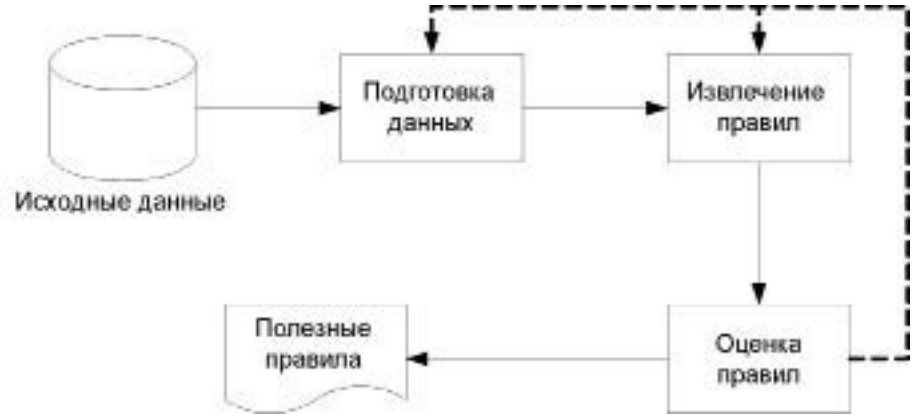


Рисунок 2 – Этапы интеллектуального анализа данных

Подготовка данных

Процесс подготовки данных должен определить и обработать добываемые данные, чтобы сделать их пригодными для конкретных методов интеллектуального анализа. Подготовка данных является первым важным шагом на пути интеллектуального анализа и играет в нем решающую роль. Как правило, подготовка данных включает в себя четыре процесса:

1. Очистка данных. Должна заполнить вакантные значения данных, устранить зашумленные данные и исправить несогласованность в данных.

2. Выбор данных. Должен определить расположение используемых в данном анализе данных.

3. Предварительная обработка данных. Является расширением процесса очистки данных, которые были выбраны.

4. Выражение данных. Должно преобразовать данные после предварительной обработки в форму, которая может быть принята по условию алгоритма анализа данных, основанного на нейронной сети.

Анализ данных, основанный на нейронной сети, может работать только с числовыми данными, из чего следует, что необходимо преобразовывать символьные данные в числовые. Простейший способ заключается в создании таблицы соответствий между символьными данными и числовыми.

Извлечение правил

Существует множество методов извлечения правил, среди которых наиболее часто используются LRE (Limited Relative Error) метод, метод черного ящика, метод извлечения нечетких правил, метод извлечения правил из рекурсивной сети, алгоритм извлечения правил двоичного входа и выхода (BIO-RE), алгоритм частичного извлечения правил (Partial-RE) и алгоритм полного извлечения правил (Full-RE).

Правила оценки

Несмотря на то, что цель правил оценки зависит от конкретного применения, в общем они могут быть оценены в соответствии со следующими задачами:

1) найти оптимальную последовательность извлечения правил; сделав это, получим лучшие результаты в ряде определенных данных;

2) проверить точность извлеченных правил;

3) определить количество знаний в нейронной сети, которые не были извлечены;

4) определить противоречия между извлеченными правилами и обученной нейронной сетью.

Анализ различных видов нейронных сетей

Существует множество алгоритмов анализа данных, основанных на нейронных сетях, проанализируем два наиболее популярных, основанных на самоорганизующихся нейронных сетях и на нечетких сетях.

Анализ данных, основанный на самоорганизующейся нейронной сети

Самоорганизационный процесс - процесс обучения без учителя. При таком обучении обучающее множество состоит из значений входных переменных, а в процессе обучения нет сравнивания выходов нейронов с желаемыми значениями. Можно сказать, что такая сеть учится понимать структуру данных.

Идея сети Кохонена принадлежит финскому ученому Тойво Кохонену. Принцип работы этих сетей заключается во введении в правило обучения нейрона информации о его расположении, то есть составляются карты размещения нейронов.

Самоорганизующиеся карты Кохонена используются для моделирования, прогнозирования, поиска закономерностей в больших массивах данных, выявления наборов независимых признаков и сжатия информации [4, 5, 6].

Анализ данных, основанный на нечеткой нейронной сети

В основе нечетких нейронных сетей лежит идея использования существующей выборки данных для определения параметров функций принадлежности, выводы делаются на основе аппарата нечеткой логики, а для нахождения параметров функций принадлежности используются алгоритмы обучения нейронных сетей. Такие системы могут использовать заранее известную информацию, обучаться, приобретать новые знания, прогнозировать временные ряды, выполнять классификацию образов. Но одним из главных достоинств является наглядность работы такой сети для пользователя.

Из таблицы 1 видно, что и сети Кохонена, и нечеткие нейронные сети имеют достоинства и недостатки.

Таблица 1 - Преимущества и недостатки популярных нейронных сетей в data mining

Тип нейросети	Область применения	Преимущества	Неодстатки
Сеть Конохена	Классификация, кластерный анализ, прогнозирование, сжатие данных	Устойчивость к зашумленным данным, неуправляемое обучение, быстрое обучение, возможность визуализации, возможность упрощения многомерной структуры	Эвристичность алгоритма обучения, предопределенность числа кластеров
Нечеткая нейронная сеть	Прогнозирование, классификация	Хорошая сходимость, быстрое обучение, интерпретируемость накопленных знаний, наглядность работы, легко определить размер сети, допустимость к зашумленным и неточным данным, способны аппроксимировать функции любой степени нелинейности, параллельные вычисления	Априорное определение компонентов

Основное отличие сетей Кохонена от других типов нейронных сетей состоит в наглядности и удобстве использования. Эти сети позволяют упростить многомерную структуру, их можно считать одним из методов проецирования многомерного пространства в пространство с более низкой размерностью. Другое принципиальное отличие сетей Кохонена от других моделей нейронных сетей - неуправляемое или неконтролируемое обучение, что позволяет задавать лишь значения входных переменных. Важнейшим преимуществом нейронечеткой сети является возможность построения одной сети для вычисления нескольких выходных значений по нескольким входным, а также способность к логическому описанию процессов и ручной корректировке функций принадлежности.

Однако нечеткие нейронные сети выгодно отличаются от других типов тем, что вобрали в себя все плюсы нечетких множеств. Таким образом, объединив нечеткие множества и нейронные сети, получили универсальные системы, компенсирующие недостатки нейронных сетей. Основным достоинством применения нейронных сетей является возможность решать различные неформализованные задачи. При этом можно очень просто моделировать различные ситуации, подавая на вход сети различные данные и оценивая выдаваемый сетью результат.

Из рассмотренных моделей анализа данных, основанных на нейронных сетях, можно сказать, что нейронные сети, системы нечеткой логики являются прогрессивным инструментом интеллектуального поиска и извлечения знаний, т. к. обладают способностью выявления значимых признаков и скрытых закономерностей в анализируемых показателях

Реализация анализа журнала действий пользователя для определения наиболее частых действий на основе нейросети Кохонена

Для решения конкретной задачи – анализа данных о действиях пользователей ПО, был использован алгоритм ограниченного перебора, с установлением ассоциаций. Этапы этого алгоритма показаны на рис.3.

Анализ действий пользователя сводится к задаче обнаружения знаний в базах данных, называемый KDD (Knowledge Discovery in Databases). Это процесс поиска полезных знаний в 'сырых данных', который включает в себя вопросы, позволяющие обнаруживать знания.

Этими знаниями могут быть правила, описывающие связи между свойствами данных (деревья решений), часто встречающиеся шаблоны (ассоциативные правила), а также результаты классификации (нейронные сети) и кластеризации данных (карты Кохонена) и т.д.

Для решения поставленной задачи будет использован пакет Deductor – полнофункциональный инструмент для Knowledge Discovery in Databases.

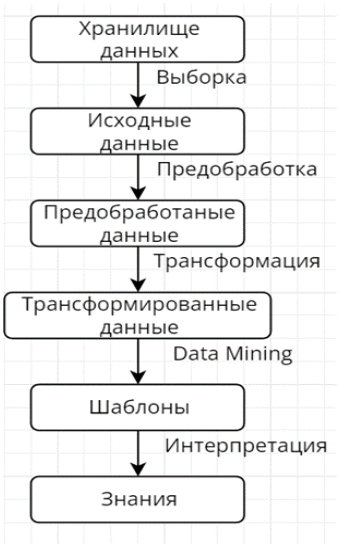


Рисунок 3 – Процесс получения знаний о действиях пользователя

Прежде всего должна быть выполнена подготовка исходного набора данных. В каждом приложении из состава пакета для этого предназначен специальный мастер подключения. Мастер позволяет импортировать данные из СУБД. Следующий шаг алгоритма - предобработка данных (удаление пиковых значений). Для выполнения этого шага в пакете существует приложение RawData Analyzer.

Далее необходимо выполнить трансформацию (нормализацию) данных. Многие приложения из состава пакета производят трансформацию данных автоматически. Например, для нейронных сетей, приложение само переводит числовые поля в нужный диапазон (нормализует), преобразует строковые, булевые и поля типа дата к числовым значениям.

После проведенной подготовки данных выполняется основной этап - Data Mining. В состав пакета включены приложения, реализующие популярные и эффективные методы DM. Neural Analyzer – нейронные сети, Tree Analyzer – деревья решений, Somap Analyzer – самоорганизующиеся карты Кохонена [7, 8, 9].

Для проверки алгоритма была рассмотрена задача анализа действий над паролями пользователя. Анализ выполнялся по шести критериям (количество изменений пароля, максимальный и минимальный период действия пароля, минимальная длина пароля и т.д.). На рис.4 показаны карты Кохонена по каждому критерию. Цветом отображена степень редкости того или иного события. За пределами двух линий находятся информационные шумы.

Кроме графической визуализации результатом алгоритма являются значения по каждому критерию, позволяющие выполнить анализ действий над паролями (табл.2).

Все приложения из состава пакета позволяют эффективно использовать полученные знания или модели на других данных [10].

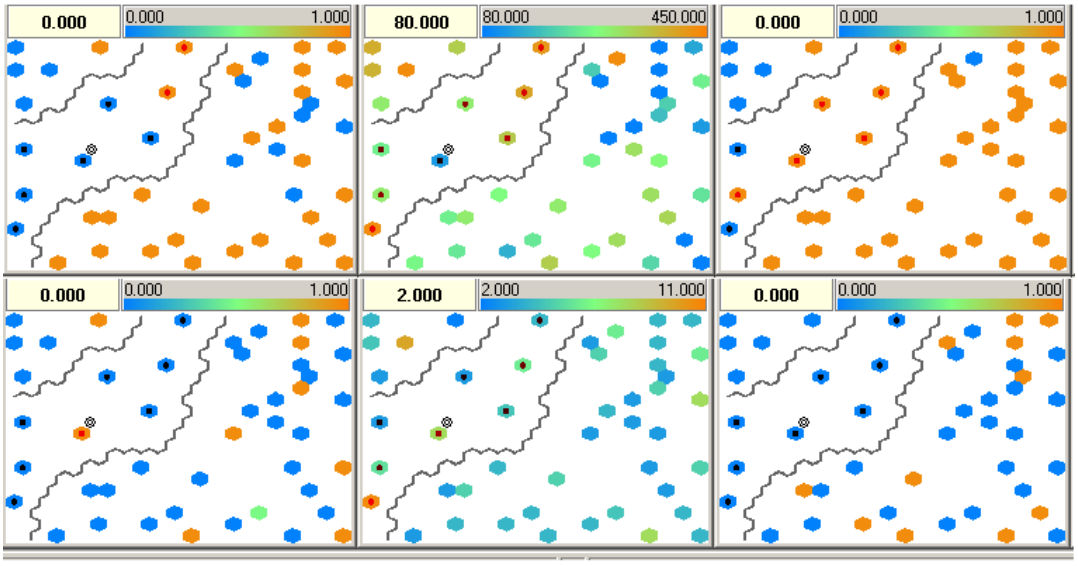


Рисунок 4 – Карты Кохонена

Таблица 2 – Результат анализа действий над паролями

Принудительное сохранение истории паролей	24 Пароля сохранено
Максимальное время действия пароля	42 дня
Минимальное время действия пароля	1 день
Минимальная длина пароля	7 символов
Соответствие требованиям сложности	Включено
Обратное шифрование в паролях	Выключено

Пакет Deductor удовлетворяет всем требованиям для успешного взаимодействия с экспертом, имеет развитый интерфейс, поддержку различных форматов хранения данных, интеграцию с офисными пакетами и т.д.

Заключение

Основные результаты работы следующие:

- использование нейронных сетей возможно для анализа действий пользователя, и имеет ряд преимуществ, но требует продуманного подхода к выбору типа нейросети;
- только средствами нейросетей на данный момент затруднительно проводить анализ данных пользователя, поэтому наиболее оптимальным вариантом является сочетание нейронных сетей, в частном случае сети Конохена, баз данных, и пакетов для предобработки;
- процесс получения знаний о действиях пользователя в условии наличия журнала записей не требует знаний в программировании, вполне можно выполнить все этапы процесса при помощи сторонних программ;
- данные о действиях пользователя не только помогут узнать на что конкретно делать упор при разработке приложения, но и позволять создателю ПО модернизировать уже готовые версии под новые потребности пользователя не нагромождая программу.

Литература

1. Дюк, В.А. Data Mining - интеллектуальный анализ данных // Информационные технологии: сайт. - URL: http://www.inftech.webservis.ru/it/database/datamining/ar2.html (дата обращения 01.11.2010)

2. Манжула, В.Г. Методы «мягких» вычислений для аналитической обработки информации в условиях неопределенности / В.Г. Манжула, С.А. Морозов, С.В. Федосеев // Фундаментальные исследования. - 2009. - № 4. - С. 75-76.

3. Назаров, А.В. Нейросетевые алгоритмы прогнозирования и оптимизации систем/ А. В. Назаров, А. И. Лоскутов - СПб.: Наука и Техника, 2003. - 384 с.

4. Чубукова, И. А. Data Mining. - М.: Изд-во «Интернет-университет информационных технологий - ИНТУИТ.ру», 2006. - 384 с.

5. Ярушкина, Н. Г. Основы теории нечетких и гибридных систем: учеб. пособие. - М.: Финансы и статистика, 2004. - 320 с.

6. Xianjun, Ni. Research of Data Mining Based on Neural Networks // World Academy of Science, Engineering and Technology. - 2008. - № 39. - P. 381-384.

7. Редько, В. Г. Эволюция, нейронные сети, интеллект: Модели и концепции эволюционной кибернетики / В. Г. Редько. - М.: Ленанд, 2015. - 224 с.

8. Редько, В. Г. Подходы к моделированию мышления / В.Г. Редько. - М.: Ленанд, 2014. - 392 с.

9. Яхъяева, Г.Э. Нечеткие множества и нейронные сети: Учебное пособие / Г.Э. Яхъяева. - М.: БИНОМ. ЛЗ, ИНТУИТ.РУ, 2012. - 316 с.

10. Шеннон, К. Работы по теории информации и кибернетике. М.: Иностранная литература, 1963. — 832 с.

Личман А.А., Чередникова О.Ю. Применение методов анализа данных для определения наиболее популярных функций приложения по журналу действий пользователя. *Рассмотрены методы решения задачи определения наиболее часто используемых функций приложения. Предложено использовать для этой цели нейронные сети. Выполнен анализ существующих нейронных сетей и особенностей их применения для различных целей. Предложен и реализован метод определения наиболее часто используемых функций приложения на основе нейронной сети Кохонена и прикладного пакета Deductor для предобработки данных. Это позволит разработчикам программных продуктов модернизировать уже готовые версии под новые потребности.*

Ключевые слова: *data mining, нейронные сети, сеть Кохонена*

Lichman A.A., Cherednikova O. Yu. Application of data analysis methods to determine the most popular application functions based on the user activity log. *Methods of solving the problem of determining the most frequently used application functions are considered. It is proposed to use neural networks for this purpose. The analysis of existing neural networks and the features of their application for various purposes is carried out. A method for determining the most frequently used application functions based on the Kohonen neural network and the Deductor application package for data preprocessing is proposed and implemented. This will allow software developers to upgrade ready-made versions to meet new needs.*

Key-words*: data mining, neural networks, Kohonen network*

Статья поступила в редакцию 04.05.2023

Рекомендована к публикации профессором Мальчевой Р. В.

УДК 004.946

Реализация вычислительного метода синтеза моделей трехмерных объектов по их изображениям в виде комплекса программ для решения задач виртуальной реконструкции

М.П. Руденко
Донецкий национальный технический университет
e-mail: m.p.rudenko@mail.ru

Аннотация
В статье приведена реализация вычислительного метода синтеза моделей трехмерных объектов по их изображению в виде программного комплекса для решения задач виртуальной реконструкции утраченных памятников архитектуры. Проведена проверка адекватности комплекса при моделировании эталонного объекта, а также сделано сравнение его работы с одним из популярных фотограмметрических редакторов, показавшее такие преимущества разработанного программного комплекса как полное управление процессом генерации экспертом, лучшие качественные характеристики полученной модели, небольшое время генерации, возможность генерации с использованием одного изображения, соответствие натурным измерениям эталонного объекта.

Введение

В настоящее время процесс восстановления трехмерной модели по изображениям является актуальной задачей компьютерного зрения. Активное использование трехмерной реконструкции с использованием иконографического материала в таких областях как архитектурное проектирование и реконструкция, археология, виртуальные музеи и т.д., подтверждает необходимость поисков более эффективных методов синтеза моделей трехмерных объектов по их изображениям, которые позволят сократить временные затраты при моделировании качественной модели.

Постановка задачи

Ниша компьютерных средств синтеза моделей трехмерных объектов по их изображениям, при малом количестве иконографического материала, до сих пор не заполнена в полной мере. Несмотря на многообразие методов и алгоритмов реконструкции, существующие средства, способные воссоздать трехмерную модель по достаточному числу фотографий без участия человека, генерируют модели, содержащие шумы и требующие, в связи с этим, их уточнения для обеспечения геометрической адекватности.

При решении прикладных задач моделирования в архитектуре и археологии с целью виртуальной реконструкции по фотоизображениям, являющимся единственным источником информации об утраченном архитектурном сооружении, возникает необходимость в разработке алгоритмов реконструкции, которые бы позволили получить информацию об архитектурном сооружении даже при наличии единственного изображения с использованием дополнительной информации об объекте и учетом преимущественно простой формы геометрии объекта.

Известные существующие методы не в полной мере отвечают требованиям синтеза моделей трехмерных объектов по их изображениям, таким как качество получаемой модели, скорость генерации, возможность работы с ограниченным иконографическим материалом, возможность построения модели сложной формы [1-3]. Это обусловливает необходимость совершенствования методов и алгоритмического аппарата решения задачи реконструкции моделей трехмерных объектов по их изображениям, особенно для частных случаев общей задачи с учетом объективных ограничений [4,5].

На основе обзора методов, рассмотренных в таблице 1, аргументирована необходимость в разработке вычислительного метода и программного комплекса синтеза моделей трехмерных объектов по их изображениям, с целью обеспечения высокой точности реконструкции при ограниченном количестве иконографического материала и даже с использованием одного изображения, возможности построения моделей сложной формы, а также умеренной потребности в вычислительных ресурсах.

Таблица 1. Преимущества и недостатки методов синтеза моделей трехмерных объектов по их изображениям

Методы	Преимущества	Недостатки
ВЕРОЯТНОСТНЫЙ МЕТОД РЕКОНСТРУКЦИИ	1. Автоматическая идентификация образа; 2. Возможность использования ограниченного количества изображений для распознавания.	1.Автоматическая идентификация образа с применением методов обнаружения контуров затрудняется наличием шумовых и оптических эффектов, текстурированного фона, взаимного перекрытия объектов; 2. Вероятностный подход при трехмерной реконструкции подразумевает использование определенной библиотеки моделей базовых примитивов.
МЕТОД РЕКОНСТРУКЦИИ С ПРИМЕНЕНИЕМ НАБОРА ИЗОБРАЖЕНИЙ	1. Автоматическая и полуавтоматическая идентификация образа.	1.Использование набора фотоизображений для реконструкции, что не подходит для реконструкции по одиночным изображениям; 2. Поиск удачного значения матрицы вращения и фокусного расстояния, что требует ряд итераций; 3. Отсутствие верхней границы времени, необходимого для вычисления параметров модели; 4.Точная модель может быть определена с некоторой вероятностью, которая становится больше, чем больше сделано итераций.
МЕТОД ПРОЕКТИВНОЙ ГЕОМЕТРИИ	1. Возможность использования ограниченного количества изображений для распознавания; 2. Каркасная модель получается более точной и правильной по сравнению облачной и полигональной моделью;	1. Идентификация образа предполагает участие эксперта. 2. Известные решения не позволяют качественно реконструировать кривые линии и поверхности.

Вычислительный метод синтеза моделей трехмерных объектов по их изображениям

Исходя из общей постановки задачи, разработка вычислительного алгоритма синтеза моделей трехмерных объектов по их изображениям, основанного на методе перспективных масштабов, позволяет создавать каркасную модель объекта без искажений с правильными пропорциями и отсутствием шумов с использованием даже одного фотоизображения с использованием дополнительной информации.

Последовательность синтеза моделей трехмерных масштабов, основанного на методе перспективных масштабов, включает ряд промежуточных этапов:

- определение точек схода для реконструкции архитектурного сооружения, изображенного на фотографии (одна точка схода, две точки схода, три точки схода);
- определение точки зрения для отыскания относительных натуральных величин перспективных линий архитектурного сооружения;
- отыскание относительных натуральных величин перспективных линий архитектурного сооружения, изображенного на фотографии;
- построение трехмерной модели архитектурного сооружения по найденным относительным натуральным величинам.

Входными данными служит одна фотография архитектурного сооружения, на которой с помощью эксперта наносятся опорные точки для дальнейших построений и вычислений.

Промежуточные этапы построения математической модели процесса, определяющего относительные натуральные величины архитектурного сооружения, изображенного на фотографии, по одной, двум и трем точкам схода описаны в [6,7].

Для того, чтобы избежать искажений при

построении кривых линий и поверхностей, в алгоритме синтеза моделей трехмерных объектов по их изображениям используется интерполяционный метод, описанный в [8].

Математическая модель процесса, определяющего относительную натуральную величину кривизны архитектурного сооружения, изображенного на фотографии, состоит из следующих промежуточных этапов:

– выбор способа отыскания относительной натуральной величины общего блока здания с учетом количества точек схода;
– применение численного метода при построении кривой формообразующей линии;
– сравнение результатов построения с применением и без применения численного метода.

Входными данными служит одна фотография архитектурного сооружения, на которой с помощью эксперта наносятся опорные точки для дальнейших построений и вычислений.

Для приближения искомой кривой был выбран метод интерполяции при помощи многочлена Лагранжа. Использование интерполяционного многочлена Лагранжа дает возможность построения математической модели криволинейных элементов при решении поставленной задачи независимо от среды компьютерной реализации алгоритма. Точность работы алгоритма зависит от правильности выбора опорных точек элементов реконструируемого объекта, т.е. от правильности указания соответствующих пикселей на изображении.

Доказано, что предложенный вычислительный метод синтеза моделей трехмерных объектов по их изображениям с использованием метода перспективных масштабов дает относительно простой, но качественный практический инструмент решения поставленной задачи в условиях ограниченного объема входных данных [9].

Программная реализация вычислительного метода

Реализация вычислительного алгоритма производится с помощью разработанного комплекса программ, включающего:

– программу на языке AutoLISP, встроенном в среду AutoCAD, и позволяющем разработать программу для выполнения любых геометрических построений;
– программный модуль в виде набора формул в Openoffice Calc для расчетов координат точек при построении кривых линий в качестве формообразующих элементов.

Полученная в AutoCAD модель может сохраняться для дальнейшей детализации и визуализации в других графических средах (3dMax, ArchiCAD, Cinema 4D и тд.).

Файлы расчетов с расширением .sxc открываются и обрабатываются в среде AutoCAD, что позволяет отображать результаты расчетов.

Входной информацией для создания трехмерной модели является графический файл, содержащий фотографию или другое изображение архитектурного сооружения, импортированные в рабочую среду AutoCAD. После чего на нее накладываются базовые точки для начала работы алгоритма.

Сам алгоритм состоит из следующих этапов, изображенных на рисунке 1.

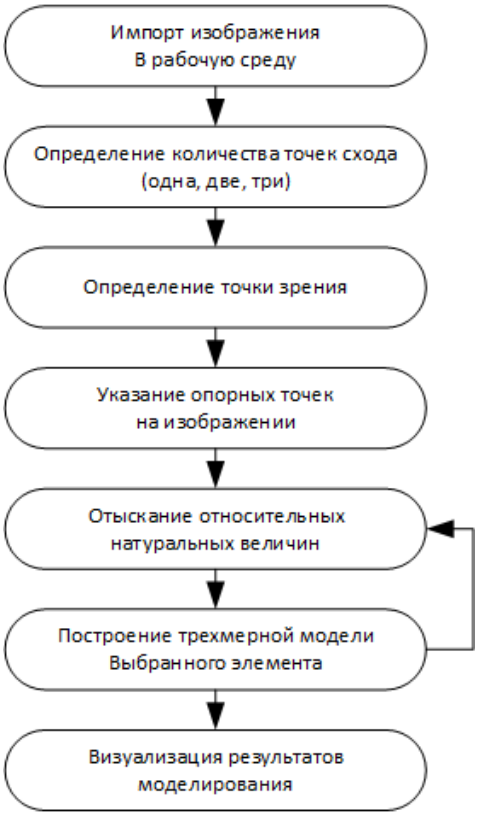


Рисунок 1 – Схема алгоритма программы

При отыскании точек схода и опорных точек элементов модели действия выполняются в графическом диалоге, точность реконструкции трехмерной модели зависит от правильности указания этих точек.

Для первичного построения трехмерной модели архитектурного сооружения, с учетом имеющейся дополнительной информации об объекте, одного изображения может оказаться достаточно. Для дальнейшего уточнения модели возможен импорт других изображений, при их наличии.

Далее работа алгоритма выполняется с помощью встроенного языка AutoLISP. Была создана программа под названием «Recon», которая запускается нажатием специальной кнопки, разработанной специально для среды AutoCAD (рис. 2а).

Выбор варианта отыскания точек схода зависит от положения архитектурного сооружения. На этапе подсказки выбора одной, двух или трех точек схода отображается диалоговое окно (рис. 2б).

Для того, чтобы пользователь определился с выбором точек схода, необходимо сравнить импортированную фотографию с изображением подсказкой.

После выбора необходимого количества точек схода, пользователю предлагается нанести опорные точки на фотографию. В рабочем окне появится подсказка, откуда следует начинать наносить опорные точки. После нанесения опорных точек, программа выполняет построение выбранной части здания с вопросом о выборе базовой точки, а также с указанием вида базовой модели для первичного построения (рис. 2в).

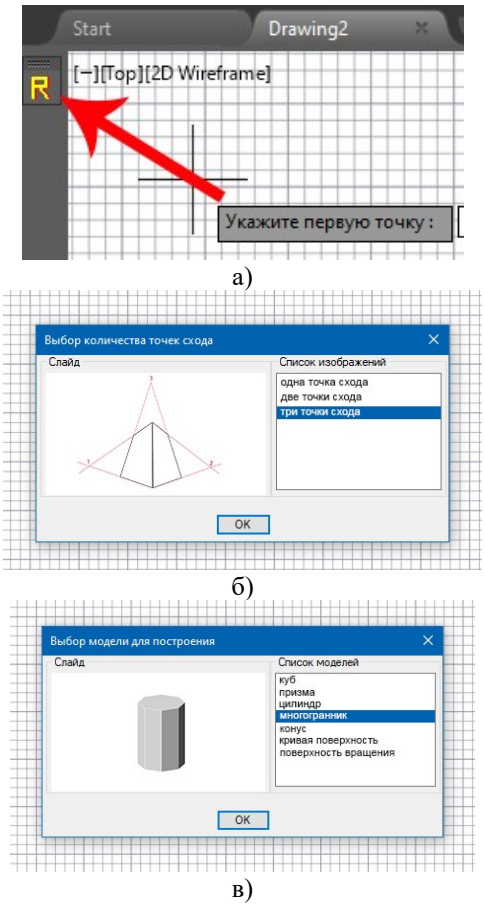


Рисунок 2 – Примеры: а) отображение иконки модуля «Recon»; б) окно выбора количества точек схода; в) окно выбора модели для построения

Для компьютерного моделирования объектов, в которых кривая линия выступает формообразующем элементом, разработан плагин, обеспечивающий совместную работу AutoCAD и Openoffice Calc с помощью плагина OMA_dwg. Так как программа должна не только открыть книгу Openoffice Calc, но еще и обратиться к нужному листу, то в листе «Расчеты» для облегчения программирования введены дополнительные данные, которые уточняют характеристики, создаваемой в рисунке AutoCAD кривой. После формирования table_items программа ex_break_connect закрывает соединение с табличным процессором Openoffice Calc и выгружает его из памяти. При этом проверяется, какие глобальные переменные, содержащие указатели объектов Openoffice Calc, сформированы, и их объекты аннулируются с помощью функции vlax-releaseobject.

Так как в рабочей среде редактора AutoCAD моделирование изначально строится на геометрических примитивах, то архитектурные детали следует назначать в них. Например: крыша и карниз – призма, прямоугольный вид здания – параллелепипед, башня – цилиндр или N-угольная призма, крыша на башне – конус или поверхность вращения, кривой фасад здания – цилиндрическая поверхность.

Команда построения трехмерной модели будет работать до тех пор, пока будет запрос на построение.

Таким образом, трехмерная модель здания будет состоять из отдельных блоков, которые легче поддаются редактированию по сравнению со слитными моделями.

Проверка адекватности

Проведено испытание программного комплекса для моделирования эталонного объекта, а также было сделано сравнение полученной модели эталонного объекта с моделью, сгенерированной в фотограмметрическом редакторе Autodesk ReCap Photo. Данный редактор лучше всего справляется с задачей генерации трехмерной модели объекта по его фотографиям.

В результате эксперимента можно сделать вывод о том, что модель, полученная с помощью разработанного программного комплекса, отличается лучшими качественными характеристиками по сравнению с моделью, построенной в Autodesk ReCap Photo, так как не содержит шумов и искажений, которые отображаются в другой модели, а, следовательно, не требует дальнейшей ее доработки после генерации в графической среде AutoCAD.

Время генерации модели в Autodesk ReCap Photo составляет один час, процесс генерации неуправляем экспертом, а результат

зависит от правильно выравненных камер на фотоизображениях. Так как выравнивание камер происходит автоматически, то хорошего результата можно добиться путем нескольких итераций. Модель на выходе получается полигональной и содержит большое количество артефактов, а значит, ее редактирование в дальнейшем может занимать до 24 часов.

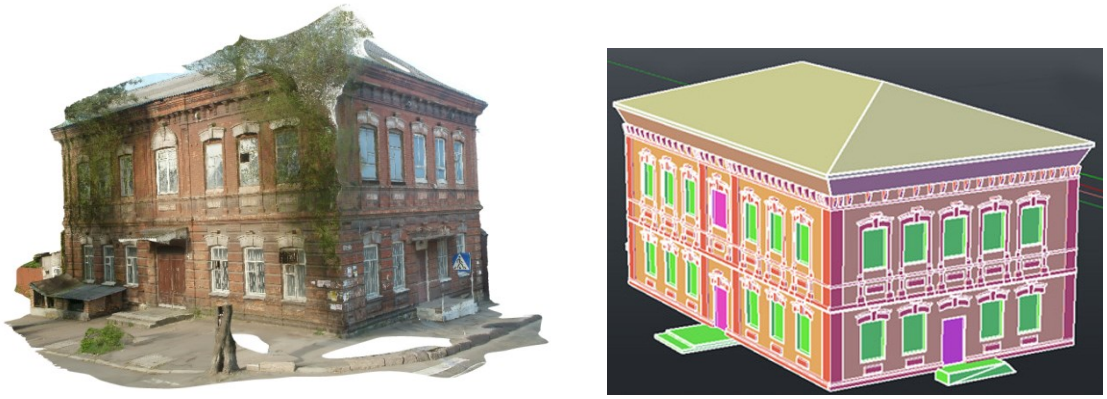
Компьютерная модель, полученная в разработанном программном комплексе, на выходе состоит из определенного количества блоков. Скорость генерации одного блока – от одной до трех минут. Количество блоков эталонного здания насчитывает семь, включая мелкие детали, такие как окна и декор над окнами. Процесс копирования деталей (окна и двери) занимает до одной минуты, а количество скопированных элементов составляет сто шесть элементов. Процесс генерации полностью управляем экспертом, а полученная модель не требует дальнейшего редактирования контуров. Сравнение общего времени реконструкции модели в редакторе Autodesk ReCap Photo и полученной с помощью разработанного программного комплекса также показано в таблице 2.

Главным преимуществом разработанного комплекса программ является возможность генерации трехмерной модели объекта, используя одно изображение, что является невозможным для редактора Autodesk ReCap Photo.

Сравнение результатов моделирования показано на рисунке 3.

Таблица 2 - Сравнение времени реконструкции модели в редакторе AutodeskReCapPhoto и полученной с помощью разработанного программного комплекса

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ	AUTODESK RECAP PHOTO	РАЗРАБОТАННЫЙ ПРОГРАММНЫЙ КОМПЛЕКС
Количество фотографий для первоначальной реконструкции (шт)	от 20ти по периметру здания	1 или 2
Время фотографирования (час)	от 20 минут (если рассчитывать 1 минуту на 1 фотографию)	-
Время генерации модели (час)	От 1 часа	От 1 до 3 минут на один блок (7 блоков) и 1 минута на копирование одного элемента с вставкой (106 элементов)
Дальнейшая доработка модели (час)	Требуется (AutoCAD или 3dsMax), до 24 часов	в AutoCADне требуется
Общее время реконструкции	26 часов	2-4 часа



а) б)

Рисунок 3 - Сравнение результатов моделирования: а) модель, сгенерированная в Autodesk ReCap Photo; б) модель, полученная с помощью разработанного программного комплекса

Модель, полученная с помощью разработанного программного комплекса, отличается лучшими качественными характеристиками, по сравнению с моделью, построенной в одной из самых популярных систем –Autodesk ReCap Photo, такими как: полное управление процессом генерации экспертом; полученная модель не требует дальнейшего редактирования контуров; время генерации модели в ряде случаев может быть сокращено до 5 раз; возможность генерации модели с использованием одного изображения; соответствие натурным измерениям эталонного объекта.

Результаты применения разработанного программного комплекса приведены при виртуальной реконструкции частично или полностью утраченных памятников архитектуры г. Донецка по сохранившимся фотоизображениям времен Старой Юзовки и подробно описаны в [10-12].

Выводы

Таким образом, выявлены следующие преимущества разработанного программного комплекса: полное управление процессом генерации экспертом; полученная модель не требует дальнейшего редактирования контуров; время генерации в ряде случаев может быть сокращено до 5 раз; возможность генерации с использованием одного изображения; соответствие натурным измерениям эталонного объекта.

Перспективным направлением дальнейшего усовершенствования программного комплекса является задача трехмерной реконструкции объектов культурного наследия, не являющимися архитектурными объектами.

Литература

1. Крейдун Ю.А. Построение пространственных моделей утраченных архитектурных памятников по одиночным изображениям /Ю.А. Крейдун, С.И. Жилин // Ползуновский вестник, № 3, 2004. – С. 83-88.

2. Меженин А.В. Реконструкция трехмерных моделей по растровым изображениям / А.В. Меженин, В.Т. Тозик // Научно-технический вестник информационных технологий, механики, оптики, № 45, 2007. – С. 203-207.

3. Захаров А.А. Трехмерная реконструкция визуальной обстановки по видеоизображениям на основе вероятностного подхода / А.А. Захаров, А.Ю. Тужилкин // Радиотехнические и телекоммуникационные системы, № 2, 2014. – С. 45-49.

4. Хапаев, В.В. Компьютерная 3D реконструкция античного и средневекового города Херсонес Таврический: опыт, проблемы и перспективы [Текст] / В.В. Хапаев, И.В. Бацура // Историческая информатика. – 2018. – № 4. – С. 39 - 56.

5. Виртуальная реконструкция историко-культурного наследия в форматах научного исследования и образовательного процесса: сб. науч. ст. [Текст] / под ред. Л.И. Бородкина, М.В. Румянцева, Р.А. Барышева. – Красноярск: Сибирский федеральный университет, 2012. – 196 с.

6. Руденко, М.П. Алгоритм трехмерного моделирования архитектурных сооружений по фотоизображению методом перспективных масштабов / М.П. Руденко // Информатика и кибернетика, Донецк, ДонНТУ, 2019. – №2(16). – С. 89-95.

7. Руденко, М.П. Алгоритм синтеза моделей трехмерных объектов по их изображениям с использованием перспективы с одной и тремя точками схода / М.П. Руденко //Проблемы искусственного интеллекта, Донецк, ГУ ИПИИ, 2020. – №2 (17). – С. 83-93.

8. Руденко, М.П. Усовершенствованный алгоритм синтеза моделей трехмерных объектов по их изображениям/ М.П. Руденко, А.А. Бабакина, В.В. Карабчевский // Проблемы искусственного интеллекта, Донецк, ГУ ИПИИ, 2020. – №1 (16). – С. 75-88.

9. Руденко, М.П. Моделирование сложных элементов архитектурных сооружений методом перспективных масштабов / М.П. Руденко // Информатика и кибернетика, Донецк, ДонНТУ, 2019. – №3(17). – С. 30-37.

10. Руденко, М.П. Применение алгоритма синтеза моделей трехмерных объектов по их изображениям при трехмерной реконструкции архитектурных сооружений / М.П. Руденко // Актуальные проблемы строительства, ЖКХ и техносферной безопасности: материалы VII Всероссийской (с международным участием) научно-технической конференции молодых исследователей, Волгоград, 20 - 25 апреля 2020 г. – М-во науки и высшего образования Рос. Федерации, Волгогр. гос. техн. ун-т. Волгоград: ВолгГТУ, 2020. – С. 384-385.

11. Руденко, М.П. Реконструкция утраченных архитектурных сооружений с использованием алгоритма синтеза моделей трехмерных объектов по их изображениям/ М.П. Руденко, В.В. Карабчевский // Материалы Международной научно-практической конференции, посвященной 90-летию СГТУ, «Геометрическое и компьютерное моделирование в подготовке специалистов для цифровой экономики» / г.Саратов, 2020. – С. 79-84.

12. Руденко, М.П. Виртуальная реконструкция утраченных памятников архитектуры с применением алгоритма синтеза моделей трехмерных объектов по их изображениям / М.П. Руденко // Искусственный интеллект: теоретические аспекты, практическое применение: материалы Донецкого международного научного круглого стола. − ГУ ИПИИ / г.Донецк, 2020. – С. 170-175.

Руденко М.П. Реализация вычислительного метода синтеза моделей трехмерных объектов по их изображениям в виде комплекса программ для решения задач виртуальной реконструкции. *В статье приведена реализация вычислительного метода синтеза моделей трехмерных объектов по их изображению в виде программного комплекса для решения задач виртуальной реконструкции утраченных памятников архитектуры. Проведена проверка адекватности комплекса при моделировании эталонного объекта, а также сделано сравнение его работы с одним из популярных фотограмметрических редакторов, показавшее такие преимущества разработанного программного комплекса как полное управление процессом генерации экспертом, лучшие качественные характеристики полученной модели, небольшое время генерации, возможность генерации с использованием одного изображения, соответствие натурным измерениям эталонного объекта.*

Ключевые слова: *трехмерное моделирование, вычислительные методы, автоматизация, программный комплекс, фотограмметрия, редакторы, синтез моделей трехмерных объектов.*

Rudenko M.P. Software implementation of the three-dimensional objects models synthesis method from their images solving virtual reconstruction problems. *The article presents software implementation of the three-dimensional objects models synthesis images from their method solving virtual reconstruction problems* of *lost architectural monuments . The adequacy of the complex was checked when modeling a reference object, and a comparison of its work with one of the popular photogrammetric editors was made, which showed such advantages of the developed software package as complete control of the generation process by an expert, better quality characteristics of the resulting model, short generation time, the possibility of generating using one images, compliance with natural measurements of the reference object.*

Keywords: *three-dimensional modeling, computational methods, automation, software package, photogrammetry, editors, the three-dimensional objects models synthesis.*

Статья поступила в редакцию 12.05.2023
Рекомендуется к публикации профессором Мальчевой Р. В.

УДК 519.254

Сравнительный анализ методов интеллектуальной обработки данных для повышения качества прогнозных моделей

О. В. Рычка
Донецкий национальный технический университет, г. Донецк
E-mail: olga_rychka@mail.ru

Аннотация
В данной статье отмечена важность предварительной обработки в анализе данных. Описаны результаты сравнительного анализа эффективности различных методов поиска аномальных значений в статистических данных и предложенного автором метода. Представлена программная реализация предложенного метода. Реализованный в работе метод обнаружения и обработки выбросов позволяет определять более точные значения различных показателей и способствует построению достоверных прогнозов.

Введение

Важным этапом анализа данных является их предварительная обработка с целью идентификации значений, которые не соответствуют модели поведения анализируемого процесса. Такие значения называют аномалиями. Одним из значимых инструментов анализа данных является регрессионный анализ – статистический метод, позволяющий выявлять соотношения между зависимой переменной и одной или несколькими независимыми переменными [1].

Основными двумя направлениями поиска аномалий является обнаружение выбросов и обнаружение новизны. В отличие от выбросов, новизна указывает на определённые изменения в системе и не является следствием ошибок в данных. В этом случае, задача заключается в своевременном обнаружении аномалий и анализе причин их появления, поскольку они могут сигнализировать о критически важных событиях. Целью такого поиска может являться обнаружение неисправности функционирования оборудования, сетевых хакерских атак, мошенничества с банковскими картами, выявление изменений в показателях здоровья человека, и т.д. Т.е. в данном случае исследователя интересуют сами аномалии, как индикаторы отклонения от нормального поведения системы и причины их возникновения. Поэтому, после обнаружения такие данные подвергаются дальнейшему анализу.

Основные причины возникновения выбросов – неточные измерения, некорректный ввод данных, выход из строя оборудования и т.д. В этом случае, после обнаружения аномалий их следует подвергнуть дальнейшей обработке – исключить из выборки или откорректировать [2]. Это позволит построить адекватную модель, наиболее точно описывающую существующую зависимость.

Чтобы репрезентативность выборки не была снижена, исключение аномалий из неё можно осуществлять, когда она содержит достаточное количество данных.

Основными методами корректировки являются:

- ручная замена выброса на другое, более подходящее значение;
- изменение экстремальных значений на наиболее вероятное значение;
- сглаживание данных;
- интерполяция аномалий. Они заменяются значениями, которые получены на основе ближайших соседей.

Целью исследования является проведение сравнительного анализа эффективности различных методов поиска аномальных значений в статистических данных, а также сравнение их с методом, предложенным автором.

Основные методы обнаружения выбросов

На сегодняшний день существуют следующие методы распознавания аномалий:

- статистический анализ;
- кластеризация;
- алгоритм ближайшего соседа;
- классификация;
- спектральные методы;
- гибридные методы.

При использовании статистического анализа определяется разница между построенной моделью и реальными данными. Если эта разница превышает определённый порог, то в данных существуют аномалии. Выделяют следующие группы методов статистического анализа:

- параметрические методы (на основе Гауссовой модели, на основе регрессионной модели, их комбинация);
- непараметрические методы (методы на основе гистограмм или функций ядра).

Кластеризация заключается в том, что все похожие экземпляры группируются в кластеры, если какой-либо экземпляр удален от центров кластеров более чем на определенную величину, то он считается аномальным. Также аномальными могут быть признаны разрозненные и незначительные кластеры.

В алгоритмах ближайшего соседа осуществляется определение расстояния или меры сходства между двумя экземплярами данных.

Метод классификации заключается в том, что наблюдения делятся на один или несколько классов, а те наблюдения, которые не принадлежат ни к одному из классов, признаются выбросами. Самыми распространенными подходами в этом методе являются:

- нейронные сети;
- Байесовы сети;
- метод на основе правил;
- метод опорных векторов.

При использовании спектрального метода на основе частотных характеристик данных строится модель, которая должна учесть большую часть изменчивости в данных [3-7].

В таблице 1 приведены примеры основных областей и решаемых задач, в которых применяется поиск выбросов, а также наиболее часто используемые методы применительно к каждой области.

Таблица 1 – Примеры применения методов поиска выбросов в различных областях

Область	Пример задачи	Метод
Медицина	вспышки заболеваний, отклонения в состоянии пациентов, ошибки записи	параметрические статистические методы, нейронные сети, байесовские нейронные сети, методы на основе правил, алгоритм ближайших соседей
Астрономия	отделение квазаров (активное ядро галактики) от звёзд	алгоритм ближайшего соседа
Компьютерные сети	обнаружение сетевых вторжений, взломов	все виды статистических методов, все виды классификации, кластеризация, ближайшего соседа
Обнаружение мошенничества	мошенничество с кредитными картами, мобильными телефонами, страховые агентства	статистические методы с использованием гистограмм, параметрические статистические методы, нейронные сети, методы на основе правил, кластеризация
Промышленность	поломки оборудования	параметрические и непараметрические статистические методы, нейронные сети, спектральный анализ
Торговля	выявление аномального спроса	параметрические статистические методы
Обработка изображений	спутниковые изображения, распознавание цифр, медицинские снимки	параметрические статистические методы, нейронные и байесовские сети, кластеризация, алгоритм ближайшего соседа

Как видно из таблицы, параметрические статистические методы используются для поиска аномальных значений в выборке практически во всех предметных областях.

Такое положение делает актуальной задачу совершенствования параметрических статистических методов поиска и обработки аномалий.

Сравнительный анализ статистических методов

Для сравнительного анализа эффективности статистических методов поиска аномалий в регрессии и предложенного автором метода были выбраны следующие методы:

- Эктона;
- Титьена-Мура-Бэкмана;
- Прескотта-Лунда;
- расстояние Кука;
- расстояние Махаланобиса.

Исходными данными для анализа является зависимость оборота розничной торговли непродовольственными товарами, млн. руб. от среднедушевого денежного дохода населения в Российской Федерации, руб./месяц. Для проверки работы методов, добавим в исходные данные четыре аномальных значения. В этом случае, значение коэффициента детерминации составляет 0,55.

При поиске аномальных значений методом Эктона было выявлено 3 подозрительных значения, как наибольшее отклонение исходных измерений от расчетных данных (e_i). После этого, по формуле (1) было рассчитано значение V, которое сравнивалось с критическим.

$$V = \frac{|e_k - \bar{e}|}{S_k}, \qquad (1)$$

где e_k – остаток предполагаемого выброса;

$\bar{e}$ – среднее по всем остаткам.

S_k – среднеквадратическое отклонение экспериментальных точек линии регрессии с учетом отбрасывания подозрительного наблюдения.

Остаток e_i с вероятностью α считается выбросом, если расчетное значение V больше критического V_α [8]. У двух выявленных подозрительных значений Y=46359 при X=29946 и Y=45644,07 при X=32285 расчетное значение V оказалось больше критического, поэтому эти значения признаются выбросами.

Далее для поиска аномального значения использовался метод Титьена-Мура-Бэкмана [9]. Он заключается в том, что по формуле (2) рассчитывается значение R_m. После этого, полученное значение R_m сравнивается с критическим значением Rα. Если полученное значение оказывается больше, то значение Y_i является выбросом.

$$Rm = \max\left|\frac{e_i}{S_i}\right|, \qquad (2)$$

где S_i – среднеквадратические отклонения остатков.

Используя формулу (2), было выявлено всего одно подозрительное значение. Величина, полученная по критерию, составила 2,71, однако она оказалась меньше критического значения равного 2,74 для уровня значимости 0,1, поэтому подозрительное значение выбросом согласно данного метода не является.

Третий метод – метод Прескотта-Лунда. С помощью данного метода по формуле (3) было получено значение R*=2,69 для подозрительного элемента, что является меньше критического равного 2,72. Следовательно, данное значение не признаётся выбросом.

$$R^* = \sqrt{n}\max\frac{|e_i|}{\sqrt{\sum_{i=1}^{n} e_i^2}}. \qquad (3)$$

Расстояние Кука представляет собой меру влияния определённых наблюдений на построенную регрессию. Для нахождения данной статистики чаще всего используется формула 4:

$$D_i = \frac{(\hat{y}_j - \hat{y}_{j(i)})}{pS_e^2 h_{ii}}, \qquad (4)$$

где $\hat{y}_j$ – ожидаемое значение регрессии (для j-го наблюдения), построенной по всей выборке;

$\hat{y}_{j(i)}$ – ожидаемое значение регрессии, построенной по выборке без i-го наблюдения;

p – число параметров модели (для линейной оно равно 2);

S_e^2 – среднеквадратическая ошибка модели, полученная при использовании всех данных;

h_{ii} – показатель влияния i-го наблюдения на коэффициенты модели. Представляет диагональные элементы матрицы проекции на пространство регрессоров $H=X(X^TX)^{-1}X^T$. Для парной линейной регрессии значение h_{ii} находится по формуле (5):

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (5)$$

Существуют различные подходы к определению выбросов с помощью расстояния Кука. Наиболее часто используется правило, что значение с расстоянием Кука D_i более 4/n (где n – количество наблюдений в выборке) считается выбросом.

Метод Кука является наиболее трудоёмким для ручного просчёта, поэтому поиск аномальных данных осуществлялся с использованием статистического пакета R. Было выявлено, что наибольшее влияние на модель оказывают наблюдения под номерами: 12, 19, 27. (рис. 1).

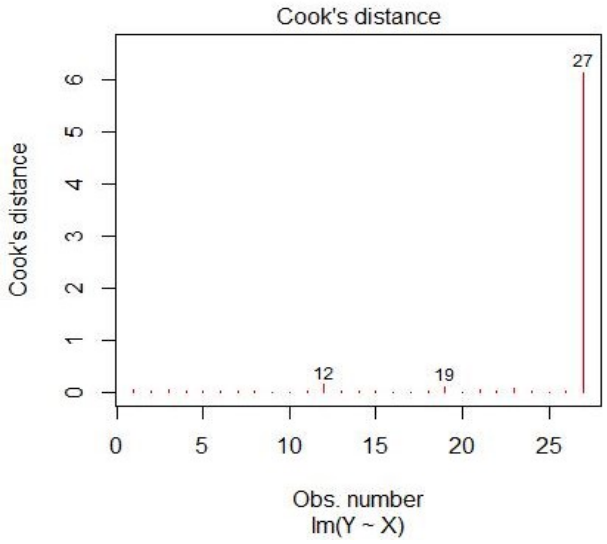


Рисунок 1 – Расстояние Кука

Эти наблюдения соответствуют следующим данным: Y=46262,16 при X=58848, Y=46359 при X=29946 и Y=45644,068 при X=32285.

Расстояние Махаланобиса определяет расстояние между двумя точками. Показывает, насколько значения наблюдений для независимых переменных отклоняются от среднего по всем наблюдениям. Расстояние Махаланобиса было рассчитано с помощью статистического пакета SPSS. В результате было найдено одно аномальное значение Y=46262,16 при X=58848.

Метод поиска аномалий, предложенный автором в [10, 11], основан на построении прямоугольной области надёжности, которая зависит от исходного уравнения регрессии. Те статистические данные, которые не попали в построенную область, признаются аномальными.

При применении данного метода были найдены все 4 аномальные измерения: Y=46262,16 при X=58848, Y=46359 при X=29946, Y=27597,16 при X=15800, Y=45644,068 при X=32285.

Результаты сравнения рассмотренных выше методов представлены в таблице 2.

Таблица 2 – Сравнительные данные методов поиска аномалий

Метод	Аномалии
Эктона	X=29946 Y=46359 X=32285 Y=45644,07
Титьена-Мура-Бэкмана	не выявлено
Прескотта-Лунда	не выявлено
Кука	X=58848 Y=46262,16 X=29946 Y=46359 X=32285 Y=45644,068
Махаланобиса	X=58848 Y=46262,16
Метод, предложенный в работе	X=58848 Y=46262,16, X=29946 Y=46359, X=15800 Y=27597,16, X=32285 Y=45644,068

Как видно из таблицы, методом поиска аномалий, предложенным автором было выявлено большее число аномалий, чем другими методами. Методом Эктона было обнаружено два аномальных наблюдения, а методами Титьена-Мура-Бэкманэ и Прескотта-

Лунда не было выявлено ни одного. Ближе всего по результативности к предложенному методу оказался метод Кука, однако данным методом были выявлены не все аномалии, а только 3, помимо этого количество элементарных операций возрастает при увеличении количества проверяемых подозрительных значений, а также отсутствует однозначный критерий того, какие из подозрительных значений признавать аномальными. Например, при использовании встроенной функции для расчёта расстояния Кука в статистическом пакете SPSS было выявлено 2, а не 3 аномальных значения. Таким образом, можно сделать вывод, что предложенный автором метод поиска аномалий является наиболее эффективным и быстродейственным.

Программная реализация предложенного метода

Для удобства использования, предложенного метода был разработан программный комплекс, которой состоит из взаимосвязанных приложений, написанных на языке C# и Visual Basic for Application для Microsoft Excel.

Пользователь может вводить данные двумя способами – вручную, непосредственно в самом приложении или загружать данные из Excel файла. Также, полученные результаты можно передать и сохранить в Excel. Вид стартового окна приложения представлен на рисунке 2. После ввода данных осуществляется их проверка на корректность и последующая сортировка.

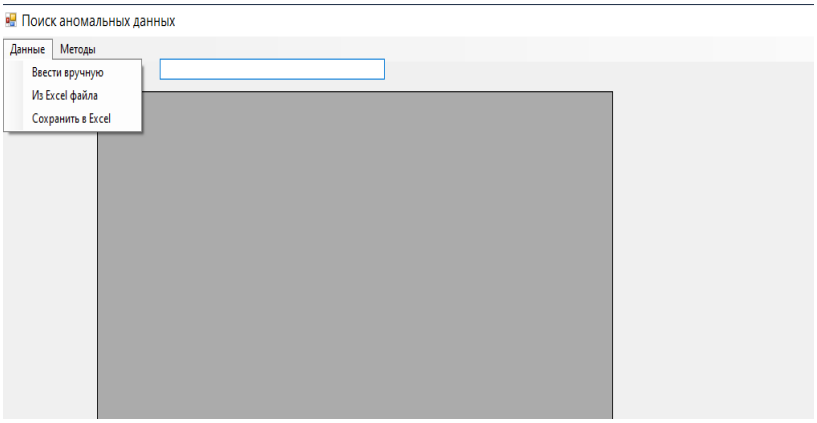


Рисунок 2 – Вид стартового окна программы

Далее пользователь может воспользоваться либо методом с последующим отбрасыванием аномальных данных, либо методом с корректировкой аномальных данных и их модификациями.

В зависимости от выбранного метода в программе рассчитываются показатели эффективности для каждого значения вероятности: коэффициенты детерминации R^2, величины доверительных интервалов, значения смещений, количество данных (исходное и после отбрасывания), точность, коэффициенты нового линейного регрессионного уравнения (рис. 3). Также, на соответствующей вкладке пользователь может посмотреть все найденные аномальные значения для различных вероятностей (рис. 4).

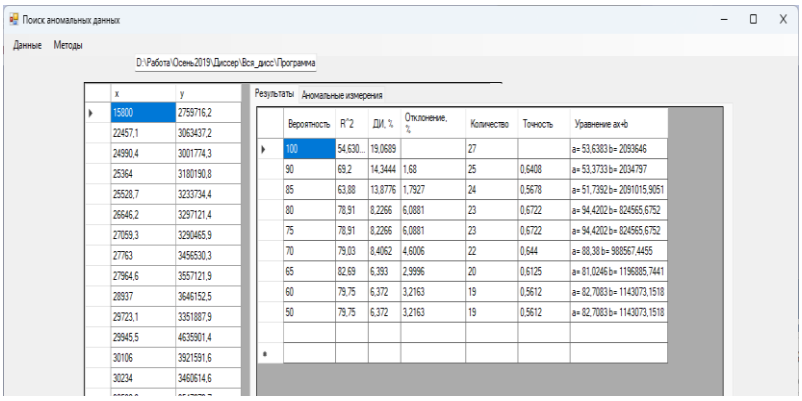
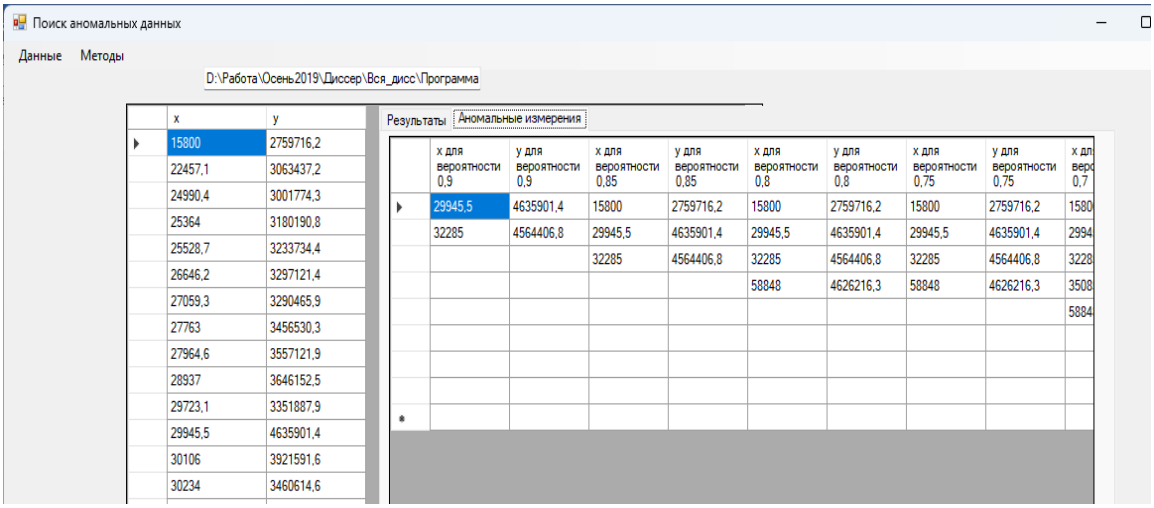


Рисунок 3 – Результаты работы программы



x	y
15800	2759716,2
22457,1	3063437,2
24990,4	3001774,3
25364	3180190,8
25528,7	3233734,4
26646,2	3297121,4
27059,3	3290465,9
27763	3456530,3
27964,6	3557121,9
28937	3646152,5
29723,1	3351887,9
29945,5	4635901,4
30106	3921591,6
30234	3460614,6

x для вероятности 0,9	y для вероятности 0,9	x для вероятности 0,85	y для вероятности 0,85	x для вероятности 0,8	y для вероятности 0,8	x для вероятности 0,75	y для вероятности 0,75	x дл вер 0,7
29945,5	4635901,4	15800	2759716,2	15800	2759716,2	15800	2759716,2	1580
32285	4564406,8	29945,5	4635901,4	29945,5	4635901,4	29945,5	4635901,4	2994
		32285	4564406,8	32285	4564406,8	32285	4564406,8	3228
				58848	4626216,3	58848	4626216,3	3508
								5884

Рисунок 4 – Аномальные данные

Разработанный программный комплекс позволяет быстро и удобно выявить аномальные данные в исходных статистических данных, устранить или откорректировать их, и получить результаты, на основе, которых можно осуществлять дальнейший анализ и прогнозирование.

Выводы

В статье рассмотрены статистические методы выявления аномальных данных, выполнен их сравнительный анализ. Помимо широкоизвестных методов в анализе использовался и разработанный автором метод. Анализ проводился на реальных данных. Наилучший результат показал метод, который был предложен в работе. С его использованием были выявлены все выбросы, содержащиеся в данных.

Для эффективного использования метода поиска аномалий, автором был разработан комплекс программ. Помимо поиска выбросов, в нём осуществляется последующая обработка выявленных аномальных измерений. Это позволяет получить адекватную модель, с использованием которой, можно строить более точные прогнозы.

Литература

1. Караулова, А.В. Применение регрессионного анализа при решении реальных задач технического характера / А. В. Караулова, И. П. Базилевский // «Молодая наука Сибири»: электрон. науч. журн. – 2020. – №3(9). - Режим доступа: http://mnv.irgups.ru/toma/39-2020.

2. Копырин, А.С. Оценка влияния аномалий на результаты анализа массивов экономических данных / А. С. Копырин, Е. В. Видищева // Modern Economy Success, 2021. - № 2. - С. 235–240.

3. Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. ACM Comput. Surv. – 41, 3, Article 15 (July 2009) – 58 pages.

4. Wilson J. Holton, Keating Barry P., Beal Mary Regression Analysis: Understanding and Building Business and Economic Models Using Excel, 2nd Edition. — New York, USA, Business Expert Press, LLC, 2016. — 205 p.

5. Кириченко, А. В. (и др.) Математические модели и методы анализа и прогнозирования: предварительная обработка результатов эксперимента, проверка статистических гипотез, корреляционный анализ, парный регрессионный анализ: учебное пособие. - Саратов: КУБиК, 2019. - 259 с.

6. Кузовлев, В.И., Орлов А.О. Метод выявления аномалий в исходных данных при построении прогнозной модели решающего дерева в системах поддержки принятия решений // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2012. № 9. URL: https://cyberleninka.ru/article/n/metod-vyyavleniya anomaliy-v-ishodnyh-dannyh-pri-postroenii-prognoznoy-modeli-reshayuschego-dereva-v-sistemah-podderzhki-prinyatiya/viewer.

7. Девянин, И.С. Предварительная обработка данных для машинного обучения // Фундаментальные и прикладные исследования в физике, химии, математике и информатике, 2021. - С. 117–121.

8. Попукайло, В.С. Обнаружение аномальных измерений при обработке данных малого объема // Технология и конструирование в электронной аппаратуре, 2016. – № 4-5 – С. 42-46.

9. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников [Текст] / А. И. Кобзарь. – М.: ФИЗМАТЛИТ, 2012. – 816 с.

10. Рычка, О.В. Разработка алгоритма реализации методов повышения качества регрессионных моделей, используемых при проектировании технических систем. // Научный журнал «Информатика и кибернетика». – Донецк: ДонНТУ, 2020. - № 3 (21). - С.42-48.

11. Рычка, О.В. Анализ эффективности усовершенствованных методов поиска и обработки аномалий для нелинейных моделей с внутренней линейностью // Международный рецензируемый научно-теоретический журнал «Проблемы искусственного интеллекта». – Донецк, 2020. – Вып. №3(18) – С. 101-110.

Рычка О.В. Сравнительный анализ методов интеллектуальной обработки данных для повышения качества прогнозных моделей*. В данной статье отмечена важность предварительной обработки в анализе данных. Описаны результаты сравнительного анализа эффективности различных методов поиска аномальных значений в статистических данных и предложенного автором метода. Представлена программная реализация предложенного метода. Анализ выполнялся на реальных данных. Реализованный в работе метод обнаружения и обработки выбросов позволяет определять более точные значения различных показателей и способствует построению достоверных прогнозов.*

Ключевые слова: *аномальные измерения, выброс, поиск аномалий, программный комплекс, сравнительный анализ, прогноз.*

Rychka O.V. Comparative analysis of intelligent data processing methods to improve the quality of predictive models. *This article highlights the importance of preprocessing in data analysis. It describes the results of a comparative analysis of the effectiveness of various methods of searching for anomalous values in statistical data and the method proposed by the author. A software implementation of the proposed method is presented. The analysis was performed on real data. The method of detection and processing of outliers implemented in the work makes it possible to determine more accurate values of various indicators and contributes to the construction of reliable forecasts.*

Key words: *anomalous measurements, outlier, search for anomalies, software package, comparative analysis, forecast.*

Статья поступила в редакцию 24.05.2023
Рекомендована к публикации профессором Зори С. А.

УДК 519.6:316.472.45

Необходимое условие оптимальности для идентификации функции активности пользователей социальной сети

М. А. Толстых, Г. В. Аверин
Донецкий государственный университет, г. Донецк
e-mail: physicisto@yandex.ru

Аннотация
Работа посвящена моделированию потоков информации в глобальных социальных сетях. Рассмотрена диффузионная логистическая модель распространения информации в социальной сети в виде одномерного нестационарного параболического уравнения, для которой поставлена задача параметрической идентификации оптимальной функции активности пользователей сети, зависящей от времени. Получено аналитическое выражение градиента целевого функционала для идентификации оптимального значения параметра-функции.

Введение

Модель диффузионной логистики является расширением статических моделей и учитывает как временные, так и структурные характеристики распространения информации в социальных сетях [1]. Её использование позволяет более точно прогнозировать распространение информации и целенаправленно воздействовать на общественное мнение. Идентификация оптимальных параметров модели диффузионной логистики является важным этапом её применения и может быть решена прямым экстремальным подходом. Оптимизация параметров позволяет получить наиболее точный прогноз диффузии информации в социальной сети и использовать его в различных целях, от бизнес-аналитики до обеспечения безопасности общества.

Рассмотрим задачу идентификации оптимального параметра, зависящего от времени, представляющего собой коэффициент активности пользователей социальной сети в модели диффузионной логистики.

Постановка задачи идентификации

Существуют различные подходы к измерению и оценке количества информации [2]. Примем следующую метрику.

Пусть x – это расстояние в графе сети, измеряемое минимальным набором рёбер, по которым может быть передана информация $v(x,t)$ от узла-источника информации в виде, например, репостов какой-либо новости. Считаем, что источник информации находится в узле x_a и в момент времени $t = t_0$ генерирует информацию $v(x_a, t_0)$ в виде одной новости. Количество репостов или их плотность (количество репостов на расстоянии x отнесённое к общему количеству узлов на этом расстоянии) будет распространяться по сети. В глобальных сетях расстояние x между узлами может становится достаточно большим, чтобы сеть считать сплошной средой. При этом, распространение информации будет подчиняется законам диффузии. Здесь состояние информационных процессов можно моделировать параболическим уравнением [1]:

$$\frac{\partial v}{\partial t} - p\frac{\partial^2 v}{\partial x^2} - uhv = 0, \quad x,t \in \Omega = (x_a, x_b) \times (t_0, t_1). \tag{1}$$

Состояние модели $v(x,t) \in L_2(\overline{\Omega})$, где L_2 – евклидово пространство функций с интегрируемым квадратом, $\overline{\Omega}$ – замыкание пространственно-временного множества Ω. В данном уравнении $p(x) \in L_2(x_a, x_b)$ – коэффициент популярности информации, который влияет на степень диффузии информации извне социальной сети. Функция $u(t) \in L_2(S)$, $S = (t_0, t_1)$ – коэффициент активности пользователей, скорость роста информации за счёт пользователей, поделившихся новостью внутри сети. Функция $h(x) \in L_2(x_a, x_b)$ – пропускная способность, т.е. максимально возможное количество поделившихся новостью пользователей на расстоянии x. Все функции $p(x)$, $u(t)$ и $h(x)$ определены на Ω, но зависят только от времени или пространства.

Граничные условия для уравнения (1) заданы на концах отрезка $[x_a, x_b]$ в виде условия первого и второго рода соответственно:

$$v = 1 \text{ на } \Gamma_a = x_a \times (t_0, t_1),$$
$$\frac{\partial v}{\partial x} = 0 \text{ на } \Gamma_b = x_b \times (t_0, t_1).$$

Информация в источнике x_a всегда равна единице, ибо содержание информации в первом узле ограничивается одной новостью, а x_b – это расстояние на котором поток указанной информации исчезает.

Начальные условия зададим в виде отсутствия обсуждаемой новости в сети в момент t_0:

$$v = 0 \text{ на } \Gamma_0 = [x_a, x_b] \times t_0.$$

Функция активности пользователей социальной сети $u(t)$ является управлением для уравнения (1). Оптимальное управление $u_*(t)$ можно найти на основании некоторых наблюдений за состоянием информации в области Ω. Отклонение модельного состояния v от экспериментально наблюдаемого v_e в реальной сети можно оценить следующим критерием качества идентификации модели в виде функционала:

$$J(u) = \iint_{\Omega} (v - v_e)^2 \, dx \, dt \in E, \qquad (2)$$

где E – это евклидово пространство действительных чисел. Наилучшее значение параметра $u = u_*$ модели (1) будет доставлять минимум целевому функционалу $J(u)$. Отметим, что $J(u)$ зависит от u не явно, а через состояние модели (1). Это существенно усложняет решение оптимизационной задачи, поскольку явную зависимость J от u выразить практически невозможно.

Рассматриваемую задачу параметрической идентификации для модели (1) можно сформулировать как экстремальную задачу [3, 4]:

$$u_*(t) = \arg \min_{u \in L_2(S)} J(u),$$
$$t \in S \subset \Omega.$$

Данную задачу целесообразно решать прямым экстремальным подходом [5,6], который предполагает непосредственную минимизацию целевого функционала (2) на основе его градиента.

$$u^{k+1}(t) = u^k(t) - b^k \alpha(t) \nabla J(u^k, t),$$
$$x \in (t_0, t_1), \qquad k = 0,1, \ldots,$$

где k – номер итерации, b^k – шаговый множитель, который выбирался методом золотого сечения, $\alpha(t)$ – параметр регулирования направления спуска [5]. Именно $\alpha(t)$ может обеспечить сходимость $u^k(t) \xrightarrow[k \to \infty]{} u_*(t)$ равномерно на S. В противном случае, если α отсутствует, то для бесконечномерных задач оптимизации (оптимизация в функциональных пространствах) сходимость к оптимуму $u_*(t)$ на S может отсутствовать или потребовать бесконечно много итераций, что, по сути, одно и тоже.

Поиск градиента

Для определения градиента необходимо выписать главную линейную часть приращения целевого функционала $J(u)$, которая представляет собой первую вариацию

$$\delta J = \langle \nabla J, \delta u \rangle_{L_2(S)}, \qquad (3)$$

где скобки означают скалярное произведение в указанном пространстве. Таким образом нам необходимо проварьировать (линеаризовать) задачу (1), (2) и привести к виду (3).

Модель (1) можно представить в виде следующего дифференциального оператора $\mathbb{D}$, действующего на v:

$$\mathbb{D} \cdot = \frac{\partial \cdot}{\partial t} - p \frac{\partial^2 \cdot}{\partial x^2} - uh \cdot.$$

При этом уравнение (1) принимает вид:

$$\mathbb{D}v = 0 \in L_2(\Omega).$$

Сначала линеаризуем целевой функционал:

$$\delta J = \iint_{\Omega} 2(v - v_e) \delta v dx dt = \langle 2(v - v_e), \delta v \rangle_{L_2(\Omega)},$$
$$\delta J \in E.$$

Теперь линеаризуем уравнение (1):

$$\delta \mathbb{D} = \mathbb{V} \delta v + \mathbb{U} \delta r \in L_2(\Omega),$$

где оператор

$$\mathbb{V} \cdot = \frac{\partial \cdot}{\partial t} - p \frac{\partial^2 \cdot}{\partial x^2} - uh \cdot$$

и оператор

$$\mathbb{U} \cdot = -hv \cdot.$$

Отобразим полученное уравнение в пространство E, где существует δJ. Там две части одной линеаризованной задачи можно будет объединить. Необходимое отображение можно сделать произвольным линейным функционалом (сопряжённая переменная) $\tilde{f} \in L_2(\Omega)$:

$$\langle \tilde{f}, \delta \mathbb{D} \rangle_{L_2(\Omega)} = \langle \tilde{f}, \mathbb{V} \delta v + \mathbb{U} \delta u \rangle_{L_2(\Omega)} \in E.$$

Преобразуем последнее выражение к такому виду чтобы в правой части скалярного произведения были δu и δv, т.е. к виду

$$\langle \mathbb{V}^* \tilde{f}, \delta v \rangle_{L_2(\Omega)} + \langle \mathbb{U}^* \tilde{f}, \delta u \rangle_{L_2(S)} \in E,$$

где звёздочка означает принадлежность оператора сопряжённому пространству.

Первое слагаемое примет вид:

$$\begin{aligned}
&\langle \tilde{f}, \mathbb{V}\delta v \rangle_{L_2(\Omega)} = \\
&= \iint\limits_{x_a t_0}^{x_b t_1} \left(\tilde{f} \frac{\partial \delta v}{\partial t} - \tilde{f} p \frac{\partial^2 \delta v}{\partial x^2} - uh\tilde{f}\delta v) \right) dxdt \\
&= \\
&= \iint\limits_{x_a t_0}^{x_b t_1} \left(-\frac{\partial \tilde{f}}{\partial t} - p \frac{\partial^2 \tilde{f}}{\partial x^2} - uh\tilde{f} \right) \delta v dxdt + \\
&+ \int\limits_{x_a}^{x_b} \tilde{f}\delta v \Big|_{t_0}^{t_1} dx + \int\limits_{t_0}^{t_1} p \frac{\partial \tilde{f}}{\partial x} \delta v \Big|_{x_a}^{x_b} dt - \\
&- \int\limits_{t_0}^{t_1} p\tilde{f} \frac{\partial \delta v}{\partial x} \Big|_{x_a}^{x_b} dt = \langle \mathbb{V}^* \tilde{f}, \delta v \rangle_{L_2(\Omega)} + \\
&+ \langle \overline{\mathbb{V}}^* \tilde{f}, \delta v \rangle_{V^*(\Gamma_0 \cup \Gamma_1 \cup \Gamma_a \cup \Gamma_b)} + \\
&+ \langle \overline{\overline{\mathbb{V}}}^* \tilde{f}, \frac{\partial \delta v}{\partial \tau} \rangle_{V^*(\Gamma_a \cup \Gamma_b)}.
\end{aligned}$$

Как мы видим, появились граничные слагаемые при δv и $\frac{\partial \delta v}{\partial \tau}$, которые отмечены соответственно операторами $\overline{\mathbb{V}}^*$ и $\overline{\overline{\mathbb{V}}}^*$.

Второе слагаемое

$$\begin{aligned}
\langle \tilde{f}, \mathbb{U}\delta u \rangle_{L_2(\Omega)} &= - \iint\limits_{x_a t_0}^{x_b t_1} h v \tilde{f} dt \delta u dx = \\
&= \langle \mathbb{U}^* \tilde{f}, \delta u \rangle_{L_2(\mathrm{S})}
\end{aligned}$$

Получаем сопряжённые операторы, действующие на $\tilde{f}$ из Ω:

$$\mathbb{V}^* \cdot = -\frac{\partial \cdot}{\partial t} - p \frac{\partial^2 \cdot}{\partial x^2} - uh \cdot, \qquad L_2(\Omega) \to L_2(\Omega),$$

$$\mathbb{U}^* \cdot = -\int\limits_{x_a}^{x_b} hv \cdot dx, \qquad L_2(\Omega) \to L_2(S).$$

Теперь можно в E объединить линеаризованные элементы задачи, пока без учёта краевых членов:

$$\delta J = \langle \mathbb{V}^* \tilde{f} + 2(v - v_e), \delta v \rangle_{L_2(\Omega)} + \langle \mathbb{U}^* \tilde{f}, \delta u \rangle_{L_2(S)}.$$

Обратимся к краевым условиям задачи (1). Проварьируем их:

$$\delta v = 0 \ \text{ на } \Gamma_a, \qquad \frac{\partial \delta v}{\partial x} = 0 \ \text{ на } \Gamma_b,$$
$$\delta v = 0 \ \text{ на } \Gamma_0.$$

Тогда краевые члены с операторами $\overline{\mathbb{V}}^*$ и $\overline{\overline{\mathbb{V}}}^*$ примут вид:

$$\int\limits_{x_a}^{x_b} \tilde{f}\delta v \Big|_{t_1} dx + \int\limits_{t_0}^{t_1} p \frac{\partial \tilde{f}}{\partial x} \delta v \Big|_{x_b} dt + \int\limits_{t_0}^{t_1} p\tilde{f} \frac{\partial \delta v}{\partial x} \Big|_{x_a}.$$

Зададим $\tilde{f} = f$ такое, чтобы все слагаемые с δv обратились в ноль. Получим сопряжённую задачу:

$$-\frac{\partial f}{\partial t} - p \frac{\partial^2 f}{\partial x^2} - fuh + 2(v - v_e) = 0, \qquad x, t \in \Omega, \tag{4}$$

с граничными и терминальным условиями:

$$pf = 0 \ \text{ на } \Gamma_a, \qquad p \frac{\partial f}{\partial x} = 0 \ \text{ на } \Gamma_b, \qquad f = 0 \ \text{ на } \Gamma_1. \tag{5}$$

Таким образом первая вариация целевого функционала принимает вид:

$$\delta J = \langle \mathbb{U}^* f, \delta u \rangle_{L_2(S)}. \tag{6}$$

Откуда следует, что градиент

$$\nabla J = -\int\limits_{x_a}^{x_b} hvf dx \in L_2(S), \tag{7}$$

который находится через решение f сопряжённой задачи (4), (6).

Необходимое условие оптимальности

Равенство нулю перовой вариации функционала $J(u)$ является необходимым условием оптимальности в задаче управления без ограничений, т.е.,

если $u = u_*$, тогда необходимо
$$\delta J(u) = 0.$$

При наличии сопряжённой задачи (4) с краевыми условиями (5) первая вариация δJ принимает вид (6) с градиентом (7), что придаёт условию оптимальности форму классического необходимого условия для экстремальной задачи:

если $u = u_*$, тогда необходимо
$$\|\nabla J\|_{L_2(S)} = 0.$$

Сравнив операторы $\mathbb{V}$ и $\mathbb{V}^*$ исходной и сопряжённой задач, мы видим, что обе задачи имеют параболический тип. Особенность сопряжённой задачи заключатся в направлении её решения во времени. Здесь время, согласно терминальному условию, является обратным, что согласуется со знаком минус при производной $-\frac{\partial f}{\partial t}$.

Выводы

Для диффузионной модели распространения информации в социальных сетях поставлена задача параметрической идентификации как экстремальная задача для поиска коэффициентов-функций дифференциального уравнения в частных производных параболического типа.

Получено аналитическое выражение градиента целевого функционала для идентификации оптимального значения функции активности пользователей сети. Применение градиентных алгоритмов оптимизации с равномерной функциональной сходимостью обеспечит выполнение необходимого условия оптимальности в виде равенства нулю нормы градиента с требуемой точностью.

Литература

1. Wang H., Wang F., Xu K. Modeling information diffusion in online social networks with partial differential equations. Springer Nature, 2020.

2. Аверин, Г.В. О количественных величинах, характеризующих понятие «информация» // Материалы VII Международной научной конференции, посвящённой 85-летию Донецкого национального университета, 2022.– Т. 2. - С. 216

3. Толстых, В.К., Толстых М.А. Необходимое условие оптимальности параметрической идентификации для распределённой модели социальных сетей // Вестник Донецкого национального университета. Серия Г. Технические науки. – 2021. - №3. – С. 63-68.

4. Толстых, М.А. Задача идентификации параметров социальных сетей // Материалы Международного молодежного научного форума «Ломоносов-2020». - М.: МАКС Пресс, 2020. (URL: http://lomonosov-msu.ru/archive/Lomonosov_2020/data/section_34_19485.htm)

5. Толстых, В.К. Прямой экстремальный подход для оптимизации систем с распределёнными параметрами. - Донецк: Юго-Восток, 1997.

6. Васильев, Ф.П. Методы оптимизации. Т. 2. - М.: МЦНМО, 2011.

Толстых М. А., Аверин Г. В. Необходимое условие оптимальности для идентификации функции активности пользователей социальной сети. *Работа посвящена моделированию потоков информации в глобальных социальных сетях. Рассмотрена диффузионная логистическая модель распространения информации в социальной сети в виде одномерного нестационарного параболического уравнения, для которой поставлена задача параметрической идентификации оптимальной функции активности пользователей сети, зависящей от времени. Получено аналитическое выражение градиента целевого функционала для идентификации оптимального значения параметра-функции.*

Ключевые слова: *параметрическая идентификация, математическая модель, социальные сети.*

Tolstykh M. A., Averin G. V. A necessary optimality condition for identifying the activity function of social network users. *The work is devoted to modeling information flows in global social networks. A diffusion logistic model of information dissemination in a social network in the form of a one-dimensional unsteady parabolic equation is considered, for which the task of parametric identification of the optimal function of the activity of network users depending on time is set. An analytical expression of the gradient of the target functional is obtained to identify the optimal value of the parameter-function.*

Key words: *parametric identification, mathematical model, social networks.*

Статья поступила в редакцию 24.04.2023
Рекомендована к публикации профессором Павлышом В.Н.

УДК 004.4'242

Анализ применения редакторов онтологий с физической семантикой в педагогической деятельности вуза

Д.А. Филипишин[1], А.В. Григорьев[2], Е.И. Приходченко[3]
Донецкий национальный технический университет, г.Донецк
[1]domaco@mail.ru, [2]grigorievalvl@gmail.com, [3]88rapoport88@mail.ru

Аннотация
Рассматриваются популярные методы и средства интерактивных обучающих систем, а также способы использования редактора онтологий как средства CASE- и CALS-технологии. Проводится сравнительный анализ на примере условно доработанного программного продукта до CAD/CAE системы, для описания эффективности обучения будущих программистов и разработчиков, с целью более глубокого понимания парадигмы объектно-ориентированного программирования. Принципиальным отличием предложенного подхода является более приближенный метод использования онтологического инжиниринга в рамках объектно-ориентированного программирования.

Постановка проблемы

Выпускник высшей школы, особенно технической, должен обучаться на протяжении всех жизни, постоянно обновлять свои знания, повышать, развивать и углублять свой уровень компетенций – такое требование выставляют не только Европейские образовательные стандарты, но и Российские [1]. На их решение и нацелена технология интерактивного обучения.

Можно отметить работы ряда специалистов, занимающихся созданием обучающих систем и педагогическими инновациями, таких как Приходченко Е.И., Воробьёв Г.Г., Маштабиц Е.И., Шевченко О.И., Ярмухмедова К.Г. [1-7].

С другой стороны, существует такое направление, как онтологический инжиниринг (ОИ), который может рассматриваться как система формальных определений в конкретной предметной области и представленный в виде предлагаемого редактора онтологий с физической семантикой.

Физическая семантика, прежде всего, свойственна задачам САПР, предполагающих наличие CAD/CAM/CAE подсистем.

Онтологический инжиниринг способен упростить технологию интерактивного обучения для учащихся. Проблемой разработки в направлении онтологического инжиниринга занимаются ряд специалистов Боргест Н.М., Гаврилова Т.А., Соловьев В.Д., Доброхотов А.Л., Скобелев П.О., Гартман Н., Клещев А.С., Артемьева И.Л., Черняховская Л.Р., Токарев Д.В., Григорьев А.В. [8-15].

Таким образом, актуальной задачей является не решённая проблема использования онтологического инжиниринга в практике высшей школы.

Цель и задачи исследования

Исходя из поставленной проблемы, рассмотреть применение предлагаемого «усовершенствованного» редактора онтологий на учебных занятиях.

Провести анализ его использования как прикладной технологии (CASE технология):

— в виде графической интеллектуальной системы (CAM);

— умной базы данных (CAD);

— базы знаний для расчета сложных физических процессов и экспертных систем (CAE).

Также следует рассмотреть применение предлагаемого редактора онтологий в машиностроении, на примере САПР механического производства и роботостроения (CALS технология).

Сравнение CASE и CALS технологий

CASE-технология представляет собой совокупность методов и средств проектирования, разработки и сопровождения сложных систем программного обеспечения, поддержанную комплексом программных средств автоматизации. В настоящее время практически ни один серьезный программный продукт не осуществляется без использования CASE-средств.

Основная цель этой технологии заключается в том, чтобы отделить кодирование программного продукта от его проектирования на всех этапах разработки. В частности, CASE-средства делятся на две большие группы: средства проектирования спецификаций и структуры (не поддерживающие полный жизненный цикл программного продукта), а также средства генерации исходных текстов и

реализации интегрированного окружения поддержки полного жизненного цикла разработки программного продукта.

Чаще всего, понятие «CASE-технология» ассоциируется с UML или семейством стандартов IDEF0-14, позволяющими проектировать практически любой возможный программный продукт, который только может представить разработчик. Реже с интерактивными программами в целом.

Можно сказать, что CASE-технология является САПР программного обеспечения.

Основное отличие CALS-технологии от CASE в том, что первая используется для программирования механических станков, роботов и производства, как в военной, так и в промышленной отрасли. Это отличие не влияет на тот факт, что в любой момент программу по управлению действиями станка, устройства или робота можно перенести в заранее подготовленную программную среду. После чего протестировать процесс работы, проанализировать статистику, сделать экономические выводы и внести правки.

Применение CALS-технологии позволяет существенно сократить объёмы проектных работ, так как описания многих составных частей оборудования, машин и систем, проектировавшихся ранее, хранятся в унифицированных форматах данных сетевых серверов, доступных любому пользователю. На текущий момент считается, что успех на рынке (особенно в век развития нейронных сетей и роботостроения) сложной технической продукции будет немыслим вне технологий CALS.

Главной проблемой построения открытых распределённых автоматизированных систем для проектирования и управления в промышленности является обеспечение единообразного описания и интерпретации данных, независимо от места и времени их получения в общей системе, имеющей масштабы вплоть до глобальных. Отсюда становится реальной успешная работа над общим проектом разных коллективов, разделённых во времени и пространстве и использующих разные CAD/CAM/CAE-системы.

Онтологический инжиниринг и редактор онтологий

Определим основные используемые понятия, необходимые для освоения онтологического анализа. Онтология – философское учение об общих категориях и закономерностях бытия, существующее в единстве с теорией познания и логикой [8]. Онтология (в информатике) – это попытка всеобъемлющей и детальной формализации некоторой области знаний с помощью концептуальной модели.

Онтологический анализ – разделение реального мира на составляющие и классы объектов, определение их онтологий или же совокупности фундаментальных свойств, которые определяют их изменения и поведение.

В проектировании онтологий условно можно выделить два направления:

— весомые или тяжёлые - связаны с представлением онтологии как формальной системы, основанной на математически точных аксиомах,

— лёгкие – связаны с компьютерной лингвистикой и когнитивной наукой (система абстрактных понятий).

Редакторами или конструкторами онтологий называют инструментальные программные средства, созданные специально для проектирования, редактирования и анализа онтологий. Основная функция любого редактора онтологий состоит в поддержке процесса формализации знаний и представлении онтологии как спецификации [8]. В построении онтологии используется достаточно много языков программирования, например: Ontolingua +KIF, OKBC, XOL, OWL. Обычно языки программирования в этой области формальны, они – обобщения языков – фреймов, они кодируют декларативные знания и обычно базируются на логике.

Интерес к разработке онтологий плавно нарастает как со стороны разработчиков интеллектуальных систем, так и со стороны бизнес-аналитиков. Усилия исследователей в основном были направлены на разработку технологических инструментов и примеров. Однако разработка практических онтологий в производстве, проектировании и менеджменте, особенно ИТ-менеджменте, остаётся скорее на уровне «искусства» [4].

На сегодня требуется решить ряд задач для дальнейшего развития онтологического инжиниринга, с целью более гибкого применения его на практике. Прежде всего необходимо решить задачу внедрения в «тяжёлую» онтологию средств объектно-ориентированного программирования, после чего реализовать принципы метапрограммирования [16]. Такой подход позволит выполнять простые расчеты и управлять механизмом наследования наглядно, что может служить отличным инструментом для обучения начинающих программистов.

Решение этих двух задач не позволит решить главную задачу применения редакторов онтологий в обучении и производстве в полной мере без дополнительных инструментов. Для этого, редактор онтологий с физической семантикой потребуется каждый раз адаптировать для различных прикладных задач, согласно требуемой сигнатуре.

Рассмотрим предлагаемый редактор онтологий в обучении.

Редактор онтологий с физической семантикой как воркшоп метод в обучении

В эпоху пандемии рабочий процесс наряду с обучением переместился в дистанционный режим. В виду этого интерактивные средства дистанционного обучения получили резкий скачок в развитии и значительный прирост в посещении и использовании. Учителя получили возможность вести личные чаты с родителями, не дожидаясь пока ребёнок приведёт родителя в школу, а также вести всю или как минимум большую часть документации, касающейся учебного процесса. Эти же возможности стали доступны и педагогам в вузах, за исключением наличия курсовых проектов и работ, а также потребности хранить выполненные задания какое-то время, что требует обязательного наличия канала передачи файлов с выполненным заданием.

Программные средства, аналогичные Moodle, позволяют проводить полный жизненный цикл обеспечения обучающих курсов: создание и регистрация участников, отправка, прием, получение и проверка задания, выставление оценок, а также обратная связь со всеми участниками текущего конкретного курса в случае несогласия с выставленной отметкой, -и всё это в одной программе. Конечно же, любая программная система подвержена тенденциям изменения и устаревания, в виду чего требует наличия минимальных знаний для всех участников и навыков адаптирования к изменениям от версии к версии программного продукта (новый формат хранения данных, новый канал или протокол передачи данных, новый тип шифрования канала передачи и хранения данных и другое).

Рассмотрим вопрос применения интерактивного обучения в высшей школе.

Образование в высшей школе на современном этапе требует обновления, наполнения новыми подходами к подаче изучаемого материала. Технология воркшоп как раз и есть той инновацией, которая в корне изменит весь ход проведения занятий на любой из их форм: лекции, семинарском или практическом занятии, либо на лабораторной работе [1].

Дефиниция «воркшоп» в переводе с английского обозначает «мастерская». Суть технологии состоит в активизации всех студентов, задействованных в учебном процессе, через самостоятельную работу в целом или её отдельный фрагмент, который очень важен в целостной структуре выполняемой работы.

Технология воркшопа состоит из таких методов:

— целостная – метод ретроспекции. Использование этого метода стимулирования познавательной активности обучаемого заключается в особенности преподнесения образовательного материала (на основе ранее изученного материала);

— синектика (мозговой штурм) – метод базируется на групповом решении поставленной проблемы. Применяется широко при преподнесении нового материала, когда дано основное понятие, но ещё не раскрыты все элементы изучаемого явления;

— тематическая – кейс-метод или метод анализа ситуаций. Предназначен для получения знаний по дисциплине, истина в которых плюралистична (сотрудничество студента и преподавателя);

— портфолио – метод, демонстрирующий достижения студента, его приобретенный опыт (ориентирован на личность студента);

— деловые игры – имитация реальности путем проживания ситуаций в роли и дальнейшем применении полученного опыта;

— метод проектов – письменный метод, с целью осмысления материала с разных сторон, в разных контекстах;

— «круглый стол» - метод обучает работе в группе, взаимодействию диады «студент-студент»;

— курсовая – метод развития критического, комбинаторно-логического мышления, радиантного мышления, метод эвристического обучения [2].

Технология воркшопа – это прямой путь к созданию ситуации успеха для студентов, значительному повышению их желания учиться. Искать и использовать новые знания, уметь разбираться в новых системах, а также адаптировать под требуемую предметную область.

Пробуждая интерес к учебе, повышаем и уровень самооценки, самостоятельности в коллективном деле, формируем аналитические, творческие, коммуникативные социальные навыки, расширяем границы практических компетенций, чтобы уметь действовать в новых, необычных ситуациях, сначала учебных, а в будущем – и в жизненных [3].

Следовательно, можно предположить, что при создании интерактивных обучающих систем педагогу или его помощнику не составит проблем разложить курс по составляющим, а затем после короткой беседы радоваться эффективному обучению будущих специалистов. На деле оказывается всё не совсем настолько радужно, а частенько даже наоборот. Информация чаще всего либо вырвана из контекста, что в реальных условиях требует значительной адаптации практически обесценивая полученные знания,

либо вовсе не связана с жизненными требованиями от специалистов текущего профиля на рабочих местах.

Решением этой проблемы призваны выступить программные комплексы относительно нового поколения, основанные на накоплении знаний по необходимой предметной области или концептуальной модели.

Редактор онтологий как CAM-система

Под термином «CAM-система» понимаются как сам процесс компьютеризированной подготовки производства, так и программно-вычислительные комплексы, используемые инженерами-технологами.

При использовании рассматриваемого редактора онтологий, на учебном занятии педагогу доступна возможность значительно сэкономить время и усилия на объяснение составления сложных геометрических фигур, какого-либо составного процесса с помощью постоянно обучаемой базы знаний.

Программный продукт позволяет автоматически генерировать и моделировать необходимую информацию по запросам, аналогично экспертным системам. Например, для рассмотрения некоего сложного шва на платье, относительно самого этого платья, потребовалось бы наличие всех необходимых составляющих, иначе данные получить невозможно. Текущая, условно, CAM онтологическая система способна задать пользователю серию наводящих вопросов для генерации необходимой отсутствующей информации, по которой получить искомое решение или нарисовать графическую фигуру. Это возможно благодаря реализации И-ИЛИ деревьев [17-19]. И-ИЛИ деревья – это один из подходов, используемый для представления пространства поиска решения задач технического творчества и искусственного интеллекта. Представление в виде И-ИЛИ дерева наиболее приемлемо для задач, которые естественным образом разбиваются на взаимонезависимые задачи, например, символическое интегрирование, доказательство теорем, игровые задачи и, в частности, задачи поиска технических решений.

Таким образом, этот тип программ значительно упрощает знакомство школьников с компьютерными технологиями и математикой позволяя каждому из обучаемых самостоятельно «пощупать» каждый из этапов всего необходимого процесса, не теряя возможности поиска необходимого решения и возможности задавать формальные запросы.

Редактор онтологий как CAD-система

CAD система – это автоматизированная система, реализующая информационную технологию выполнения функций проектирования.

В последнее время всё большую популярность набирают интерактивные программы, реализованные в виде видео игр, позволяющие обучающемуся программировать каждый шаг в реальном времени, либо из представленных возможностей, либо исходя из поставленной задачи. Используя подобную программу, пользователь имеет возможность использовать «умную базу данных», которая позволяет получать данные как в формальном, так и программном виде. Например, для обучения будущих программистов навыку понимания сложных алгоритмов, можно представить персонажа в некоем игровом пространстве, где запросы типа «get X±1» для получения координаты слева или справа, можно было бы заменить на «что находится сбоку», а используя диапазон значений – «что находится рядом». Этот же процесс можно эффективно применять для аппроксимации сложных функций, с целью получения графиков, отслеживания поведения сложных систем или получения иной информации.

Редактор онтологий как CAE-система

CAE-система – это общее название для программ и программных пакетов, предназначенных для решения различных инженерных задач: расчётов, анализа и симуляции физических процессов.

Программа подобного типа позволяет педагогу разбирать за учебное время практически любой физический процесс по полочкам, для демонстрации каждой или требуемой его составляющей в рамках темы занятия. Занятие по физике стали бы значительно эффективнее в плане запоминания для любого типа мышления обучающихся, а математические доказательства приобрели бы физический смысл. Например, для расчета нагрева трубы при течении пара под давлением, течение электрического тока по проводнику и т.п.

Онтологический инжиниринг, реализованный в рамках редактора онтологий, позволил бы подойти с обеих сторон физического процесса: формального и практического. Значительно упростил написание инструкций по сборке или, например, покраске изделия составного станка, для чего проектировщику достаточно получить по запросу «синяя деталь» размеры деталей, рассчитать их площадь и вычислить, какое количество синей краски

необходимо конечному пользователю для покраски готового изделия.

Онтологический инжиниринг в машиностроении (CALS-технологии)

Описанные возможности как средства CALS-технологии значительно облегчали бы программирование любой техники человеком. Причем без уточнения специфики, будь то студенческий стенд для научной работы или реально действующая военная техника.

Устройство, программируемое, к примеру, с помощью программных средств Arduino, могло бы запрашивать инструкции по составному типу, самостоятельно размечать фигуры на металлических листах для плазморезов и выполнять подобную работу более интеллектуально. Повсеместно параметры можно было бы получать формально и на родном человеку языке, по признаку или принадлежности, что позволило бы за считанные секунды составлять инструкции по сборке или ремонту, в которых требуется подробное описание всех составных частей.

Внедрение онтологического инжиниринга для сферы роботостроения на шаг бы приблизило человечество к очередной волне дискуссии о вероятной опасности слишком «умных» механических систем, по типу роботов. Само определение «робот», которому стали бы доступны накапливаемые с помощью нейросетей знания, структурируемые в базе знаний по описанию или признаку. Вызывает опасение в виду явной потери контроля над его развитием. Любое неуместное вмешательство со стороны «учителя» может повлиять на работу нейросети механической системы, которая в виду ограничений может не дополучить необходимые знания, а отсутствие ограничений может привести к непредсказуемым последствиям.

Выводы

Разнообразие предметных областей, для которых созданы онтологии по предложенной модели, доказывает универсальность этой модели, её способность описывать отношений понятий и базовые свойства, присутствующие в любой предметной области.

Таким образом, используя редактор онтологий как воркшоп или интерактивную технологию для обучения, педагог может применить множество формальных и практических методов по своему усмотрению для более наглядной демонстрации актуальной темы занятия.

Главной первоначальной задачей предлагаемого редактора «тяжёлых» онтологий с физической семантикой является приобретение обучающимися навыков разделения поставленной задачи на составляющие её части с формальной и практической точек зрения вне зависимости от предметной области и приобретения практических навыков онтологического инжиниринга.

Принципиальным отличием предложенного подхода использования редактора онтологий является более приближенный метод использования онтологического инжиниринга в рамках объектно-ориентированного программирования.

В качестве перспективного направления данной работы отметим разработку методов и средств реализации подобных систем способных применяться в учебном процессе.

Литература

1. Приходченко, Е. И. Использование технологии воркшоп для повышения креативности будущих инженеров-педагогов / Е. И. Приходченко // журнал Информатика и кибернетика, 2020. - Вып. 2 (20) – С. 65-68

2. Приходченко, Е. И. Применение технологии воркшоп для повышения уровня динамических знаний будущих специалистов / Е. И. Приходченко // журнал Информатика и кибернетика, 2020. - Вып. 3 (21) – С. 62-65

3. Приходченко, Е. И. Технология воркшоп как динамическая система подготовки будущих специалистов / Е. И. Приходченко // журнал Информатика и кибернетика, 2020. - Вып. 4 (22) – С. 63-65

4. Шевченко, О. И. Технологии нестандартного обучения / О. И. Шевченко, М. А. Волков, В. А. Леонов // Педагогика высшей школы – Казань, 2018. – № 3 (13). – С. 17-25.

5. Воробьёв, Г. Г. Школа будущего начинается сегодня / Г. Г. Воробьев. – М.: Просвещение, 2001. – 174 с.

6. Маштабиц, Е. И. Компьютеризация обучения: проблемы и перспективы / Е. И. Маштабиц, Н. Н. Моисеев. – М., 2013.-195 с.

7. Ярмухмедова, К. Г. Обучение иностранному языку по заочной форме с применением дистанционных технологий / К. Г. Ярмухмедова // Вестник Евразийской академии имени Д. А. Кунаева. – Алматы, 2015. – № 3. – С. 152-157.

8. Онтологический редактор Fluent Editor: учебно-методическое пособие к лабораторным работам / сост.: Н. М. Боргест, А. А. Орлова. – Самара: изд-во Самарского университета, 2017. – 44 с.

9. Онтологическая инженерия, философская онтология: проблемы и перспективы совместного развития / М. В. Заковоротная // Философия, этика, религиоведение, 2013.

10. Боргест, Н. М. Онтологии проектирования от Витрувия до Виттиха / Н. М.

Боргест // Онтология проектирования, 2018. – Т.8. - №4(30). – С. 487-522

11. Гаврилова, Т. А. Визуально-аналитическое мышление и интеллект-карты в онтологическом инжиниринге / Т. А. Гаврилова, Э. В. Страхович // Онтология проектирования, 2020. – Т. 10. - №1(35). - С. 87-99

12. Боргест, Н. М. Онтология проектирования: теоретические основы: учеб. пособие. - Самара: СГАУ, 2010. - 88 с.

13. Харман, Грэм. Объектно-ориентированная онтология: новая «теория всего»: пер. с англ. / Грэм Харман. – М.: Ад Маргинем Пресс, 2021. – 272 с.

14. Григорьев, А. В. Семантика модели предметной области для интеллектуальных САПР. Научные труды Донецкого государственного университета. Серия «Информатика, кибернетика и вычислительная техника», (ИКВТ-2000) выпуск 10. – Донецк: ДонГТУ, 2000. – С.148-154.

15. Онтологии и тезаурусы: модели, инструменты, приложения: учебное пособие / Б. В. Добров, В. В. Иванов, Н. В. Лукашевич, В. Д. Соловьев. – 3-е изд. (эл.) – Москва: Национальный Открытый Университет «ИНТУИТ»; Ай Пи Ар Медиа, 2020. – 172 с.

16. Краснов, М. М. Применение метапрограммирования шаблонов C++ для решения вычислительных задач // научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017 г., г. Новороссийск). – М.: ИПМ им. М.В. Келдыша, 2017. – С. 290-305

17. Кручинин, В. В. Использование деревьев И/ИЛИ для генерации вопросов и задач // Вестник Томского государственного университета, 2004. - №284. - С. 183 – 186.

18. Зорин, Ю. А. Интерпретатор языка построения генераторов тестовых заданий на основе деревьев И/ИЛИ // Доклады Томского государственного университета систем управления и радиоэлектроники, 2013. - №1. - С. 75 – 79.

19. Зорин, Ю. А. Использование алгоритмов комбинаторной генерации при построении генераторов тестовых заданий // Дистанционное и виртуальное обучение, 2013. - №6. - С. 54 – 59.

Филипишин Д.А., Григорьев А.В., Приходченко Е.И. Анализ применения редакторов онтологий с физической семантикой в педагогической деятельности вуза. *Рассматриваются популярные методы и средства интерактивных обучающих систем, а также способы использования редактора онтологий как средства CASE- и CALS-технологии. Проводится сравнительный анализ на примере условно доработанного программного продукта до CAD/CAE системы, для описания эффективности обучения будущих программистов и разработчиков, с целью более глубокого понимания парадигмы объектно-ориентированного программирования. Принципиальным отличием предложенного подхода является более приближенный метод использования онтологического инжиниринга в рамках объектно-ориентированного программирования.*

Ключевые слова*: интерактивность, онтологии, редактор онтологий.*

Filipishin D.A., Grigoriev A.V., Prikhodchenko K.I. Analysis of the use of ontology editors with physical semantics in the pedagogical activity of the university. *Popular methods and tools of interactive learning systems are considered, as well as ways to use the ontology editor as a means of CASE- and CALS-technology. A comparative analysis is carried out on the example of a conditionally modified software product to a CAD/CAE system, to describe the effectiveness of training future programmers and developers, in order to better understand the paradigm of object-oriented programming. The principal difference of the proposed approach is a more approximate method of using ontological engineering in the framework of object-oriented programming.*

Keywords*: interactivity, ontologies, ontology editor.*

Статья поступила в редакцию 14.05.2023
Рекомендована к публикации профессором Зори С.А.

УДК 004.4'242

Формирование новых экземпляров в онтологии научной и учебно-методической информации

Е. Ю. Шклярова, С. Ю. Землянская
Донецкий национальный технический университет, г. Донецк
E-mail: zsaae07@gmail.com

Аннотация
Данная научная статья представляет исследование, посвященное разработке и применению онтологии научной и учебно-методической информации. В статье рассматриваются важные аспекты развития онтологии, представлена методика формирования новых экземпляров в существующей модели. Этот процесс позволяет расширить онтологию, улучшить процессы поиска и анализа информации. Представлен программный модуль, разработанный с использованием библиотеки RDFLib, который позволяет автоматически создавать новые экземпляры на основе данных, извлеченных из научных публикаций.

Введение

В данной научной статье рассматривается онтология научной и учебно-методической информации [1]. Онтология представляет собой формальную спецификацию понятий, классов и связей между ними, которая помогает в описании и организации знаний в семантическом виде.

Одним из важных аспектов в развитии онтологии является формирование новых экземпляров в существующей модели. Это позволяет уточнить и расширить семантическую модель предметной области, обогатить онтологию новыми концептами и связями, а также улучшить процессы поиска и анализа информации. Формирование новых экземпляров может осуществляться путем автоматического извлечения данных из научных статей, учебников и других источников. Однако существуют ограничения и проблемы, связанные с неоднозначностью и разнообразием источников данных, сложностью извлечения информации из различных типов материалов, а также необходимостью постоянного обновления и поддержки онтологии.

Цель данной статьи заключается в представлении методики формирования новых экземпляров в онтологии научной и учебно-методической информации. Рассмотрены этапы анализа существующей онтологии и добавления данных в онтологию. Представлены инструменты и библиотеки, используемые для работы с онтологическими данными, и проведено сравнение некоторых из них.

Методика, описанная в статье, предоставит практические рекомендации и примеры формирования новых экземпляров в онтологии научной и учебно-методической информации, а также поможет преодолеть возможные проблемы и сложности, связанные с этим процессом

Методика формирования новых экземпляров

Первый этап методики формирования новых экземпляров в онтологии заключается в анализе уже существующей онтологии. В ходе этого анализа определяются ее текущая структура, содержание концептов, связей и атрибутов (рис. 1). В онтологии присутствуют следующие классы в соответствии с потребностями, такими как определение статей автора, извлечение полезных данных из документов и выявление статей с определенной тематикой:

– класс «Человек» предназначен для идентификации авторов статей и связи их с соответствующими документами;

– класс «Документ» представляет собой основной объект, содержащий информацию о научных статьях. К нему могут быть привязаны различные свойства и атрибуты, такие как заголовок, ключевые слова, аннотация и т.д.;

– класс «Структурная часть документа» имеет абстрактный характер и используется только в качестве базового класса для наследования другими классами, связанными со структурой документа;

– класс «Дисциплина» представляет собой класс для описания дисциплины. Он используется для классификации статей по конкретным областям знаний;

класс «Научное мероприятие» предназначен для идентификации событий, связанных с научной деятельностью, таких как конференции, семинары и прочие мероприятия

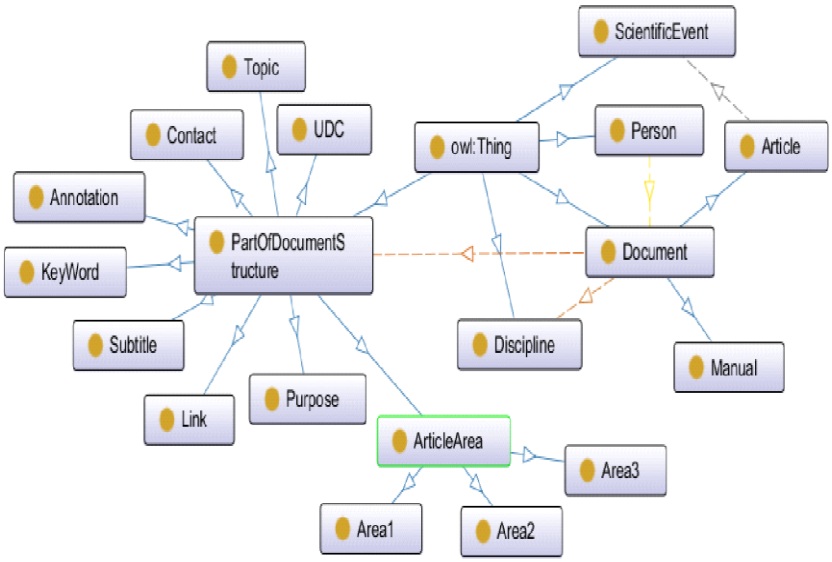


Рисунок 1 –Онтологическая модель кафедры вуза

Следующий этап включает процесс автоматического извлечения данных из научных статей. Это осуществляется с помощью методов обработки естественного языка. В процессе извлечения данных была получена информация, которая может быть использована для формирования новых экземпляров в онтологии, таких как имена авторов, УДК, список литературы, контактная информация, название статьи, аннотация, классификации и связи между понятиями.

После извлечения данных из текста статьи необходимо добавить эти данные в онтологию в виде конкретных экземпляров классов, так называемых индивидов. Каждый экземпляр обладает своими уникальными характеристиками и свойствами, которые могут быть определены в соответствии с его классом. Свойства определяют отношения и атрибуты, которыми обладает экземпляр.

Экземпляры создаются, модифицируются и удаляются в процессе работы с онтологией, что позволяет динамически расширять и изменять модель данных в соответствии с требованиями и новыми знаниями о предметной области. Для этого применяются алгоритмы сравнения и сопоставления, которые позволяют определить соответствия между извлеченными данными и существующими концептами в онтологии [2]. Это помогает установить, какие данные могут быть использованы для создания новых экземпляров или дополнения существующих концептов.

Алгоритм сравнения и сопоставления для добавления новых экземпляров в онтологию включает следующие шаги:

– создание графа для работы с онтологией и загрузка существующей онтологии;

– определение пространства имен для онтологии;

– создание нового индивида определенного класса;

– назначение уникального идентификатора и других свойств новому индивиду;

– проверка существования индивидов классов с заданными значениями;

– добавление отношений между индивидами при помощи свойств объекта;

– сохранение изменений в онтологии.

Алгоритм позволяет эффективно добавлять новые данные в онтологию и устанавливать связи между существующими и новыми индивидами, расширяя и обогащая семантическую модель предметной области.

Для работы с онтологиями можно использовать библиотеки Python, которые обеспечивают доступ к онтологическим данным и позволяют выполнять операции чтения, записи и модификации. Среди них можно выделить: RDFLib, Owlready2 и PyOWL. Проведем их сравнение

RDFLib [3] является одной из наиболее популярных библиотек для работы с RDF (Resource Description Framework) в Python. Она предоставляет функциональность для создания, чтения и записи RDF-графов, которые являются основой для представления онтологических данных. С помощью RDFLib можно устанавливать соединение с онтологией, извлекать данные, добавлять новые экземпляры и связи, а также выполнять запросы SPARQL для извлечения информации из онтологии.

Owlready2 [4] является еще одной популярной библиотекой для работы с онтологиями в Python. Она предоставляет простой и удобный интерфейс для доступа к

онтологическим данным, основанным на стандарте OWL (Web Ontology Language). Owlready2 позволяет создавать экземпляры классов, определять и изменять свойства объектов, а также выполнять запросы SPARQL-DL для извлечения информации из онтологии [5]. Библиотека также обеспечивает интеграцию с базами данных, такими как SQLite, для хранения и поиска данных.

PyOWL [6] – это Python-обертка для библиотеки OWL API, которая предоставляет функции для работы с онтологиями OWL в Python. OWL API является мощным инструментом для работы с онтологиями OWL и предоставляет поддержку различных версий OWL, таких как OWL 2 и OWL 1.1 [7].

Результаты сравнения библиотек приведены в табл. 1.

Таблица 1 – Сравнение библиотек для работы с онтологиями

Достоинства	Недостатки
RDFLib	
Широкий спектр функциональности, включая создание, чтение и запись RDF-графов, поддержку различных форматов сериализации RDF (таких как XML, JSON, N3), выполнение запросов SPARQL и тд	RDFLib имеет несколько сложный и громоздкий интерфейс, особенно для новичков в работе с RDF и онтологиями
Поддержка различных RDF-сериализаций и форматов	В некоторых случаях производительность RDFLib может быть ниже, чем у некоторых других библиотек
Хорошая документация и активное сообщество пользователей	
Owlready2	
Легкая в использовании библиотека с простым и понятным API	Могут возникать проблемы с производительностью при работе с большими онтологиями
Поддержка различных версий OWL (например, OWL 2) и возможность работы с различными типами онтологических объектов (классы, свойства, индивиды и т. д.)	Некоторые функциональности могут быть ограничены или недостаточно развиты, поскольку библиотека нацелена на простоту использования
Имеет возможность интеграции с другими библиотеками Python	
PyOWL	
Обеспечивает прямое взаимодействие с OWL API	Имеет более сложный API по сравнению с другими библиотеками, и его использование может потребовать более глубокого понимания OWL и его концепций
Обеспечивает широкий спектр функций, таких как чтение, запись, редактирование и вывод логики	

Все рассмотренные технологии являются достаточно функциональными для взаимодействия с онтологиями. Однако, учитывая несущественные недостатки, простоту использования, надежность и широкую поддержку в сообществе RDFLib, её выбор представляется предпочтительным. Эта библиотека предоставляет удобные инструменты для работы с онтологиями и позволяет эффективно обрабатывать RDF-графы.

С использованием библиотеки RDFLib был разработан программный модуль, который позволяет формировать новые экземпляры концептов онтологии, таких как научные статьи, учебники, авторы, направления и других, представленных в общей структуре онтологической модели.

Для формирования новых экземпляров модуль использует семантические правила и шаблоны, определенные в онтологии. Это позволяет системе автоматически создавать новые экземпляры на основе доступных данных и связей в онтологии.

Для рассмотренной ранее онтологии мы сможем получать следующие элементы онтологии:

– Экземпляры классов. Модуль может извлекать конкретные экземпляры людей, такие как ФИО авторов статей, конкретных дисциплин, к которым относятся статьи, а также экземпляры частей документа, таких как УДК, аннотация, название, список литературы и т.д.

– Атрибуты экземпляров: свойства или характеристики, которые присущи каждому конкретному экземпляру.

– Информация о взаимодействии и связях между различными экземплярами объектов в онтологии. Например, модуль может извлечь информацию о связи между конкретной статьей и ее авторами (свойство "isAuthor"), между статьей и ее документом (свойство "hasDocument"), а также между частями документа и самим документом (свойство "PartOfDocument").

В процессе формирования новых экземпляров модуль использует проверки и ограничения, определенные в онтологии, чтобы гарантировать соблюдение правил и структуры данных. Это поможет избежать ошибок и обеспечить согласованность информации в онтологии.

Алгоритм сравнения и сопоставления для добавления новых экземпляров в онтологию может включать следующие шаги:

Шаг 1: Импорт и чтение онтологии

– Импортирование онтологии, с которой будут сопоставляться новые данные, с использованием RDFLib.

– Чтение онтологии для получения существующих классов и свойств.

Шаг 2: Сопоставление данных с онтологией

– Итерация по каждому триплету из новых данных и проверка существования сопоставляемых концептов и свойств в онтологии.

– Если сопоставление не найдено, создание новых экземпляров и связей на основе данных из триплета.

– Если сопоставление найдено, обновление существующих экземпляров или связей в соответствии с данными из триплета.

Шаг 3: Запись изменений в онтологию

– Запись изменений и добавленных экземпляров в онтологию с использованием RDFLib.

Шаг 4: Проверка согласованности и корректности данных

– Проверка добавленных экземпляров и связей на соответствие правилам и ограничениям, заданным в онтологии.

– Если данные не соответствуют правилам, применение соответствующих корректировок или уведомление о возможных проблемах.

Рассмотрим пример формирования новых экземпляров онтологии при добавлении информации о статье «Проектирование автоматизированной системы онлайн-поиска попутчиков» [8], опубликованной в сборнике материалов конференции ИСУКМ-2021.

После извлечения данных модуль добавляет новые экземпляры в онтологию. Перечень добавленных экземпляров, сформированный в результате извлечения данных, представлен на рис. 2, а подробная информация о каждом новом экземпляре – в табл. 2. После добавления экземпляров производится проверка данных и связей для обеспечения их корректности и соответствия заданной онтологической структуре.

На рис. 3 приведена подробная информация об элементах, появившихся в онтологии при добавлении статьи.

Рисунок 2 – Новые добавленные экземпляры онтологии

Для проверки правильности результатов после добавления экземпляров в онтологию, мы можем выполнить DL-запросы (Description Logic queries) [9, 10]. На рис. 4.а показан результат выполнения запроса DL-Query для определения авторов добавленной статьи. Этот запрос помогает убедиться, что связи между статьей и ее авторами были установлены корректно.

Таблица 2 – Информация о новых экземплярах

Наимено-вание	Тип	Свойство данных	Значение свойства
udc1	UDC	udcValue	007.51
author2	Person	firstName	С
		lastName	Землянская
		middleName	Ю
author3		firstName	Е
		lastName	Мащенко
		middleName	Н
annotation1	Annotation	annotationValue	Шклярова Е. Ю., Землянская С. Ю., Мащенко Е. Н. Проектирование автоматизированной системы онлайн-поиска попутчиков. В статье обосновывается актуальность разработки автоматизированной системы онлайн-поиска попутчиков. Приведено детальное описание архитектуры разрабатываемой системы. Особое внимание уделено проектированию информационной и функциональной модели системы. Для демонстрации возможностей системы представлен прототип, базирующийся на предложенной архитектуре, рассмотрены дальнейшие направления развития системы.
key1	KeyWord	keyWordValue	система
key2	KeyWord	keyWordValue	попутчик
key3	KeyWord	keyWordValue	водитель
key4	KeyWord	keyWordValue	пользователь
key5	KeyWord	keyWordValue	поездка
key6	KeyWord	keyWordValue	поезд
link1	Link	linkValue	BlaBlaCar [Электронный ресурс] URL: https://www.blablacar.com.ua
link2	Link	linkValue	Доедем вместе! [Электронный ресур с] URL: http://www.doedemvmeste.ru
link3	Link	linkValue	Довезу! [Электронный ресурс] URL: http://www.dovezu.ru/
link4	Link	linkValue	CASE-средства для разработки информационных систем / Маклаков С.В – М.:Диалог-МИФИ, 1999.-256 с.
link5	Link	linkValue	Использование методологии моделирования IDEF при формировании структурно - параметрической модели реализации технологий обеспечения эффективного развития промышленных предприятий в условиях постиндустриальной экономики [Текст] / Тебекин А.В. // Вестник Российской таможенной академии. 2015. No 4 (33). С. 96 -103.
contact1	Contact	contactValue	zsaae07@gmail.com
area1	Area1	articleAreaValue	Информационные системы

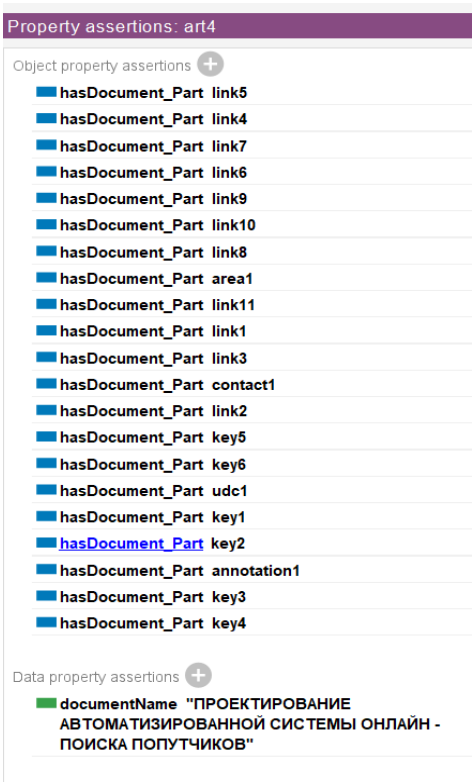
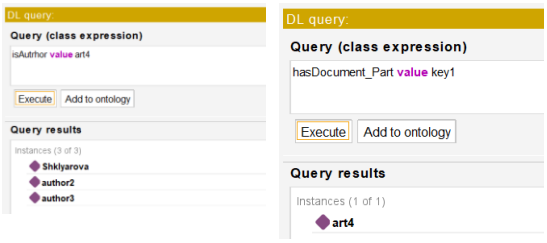


Рисунок 3 – Подробная информация о новых элементах онтологии при добавлении статьи

Далее, на рис. 4.б, представлен запрос, который проверяет правильность установления связей между ключевым словом "key1" и статьями, содержащими это ключевое слово. Результат выполнения запроса позволяет убедиться, что все соответствующие статьи были связаны правильно с ключевым словом.

Наконец, на рис. 5 выполнен запрос, который выводит всех авторов, содержащихся в онтологии. Этот запрос (Person) возвращает список URI всех объектов класса «Человек», позволяя убедиться, что все авторы были успешно добавлены в онтологию.

Результаты выполнения запросов подтверждают правильность работы алгоритмов добавления экземпляров и установления связей в онтологии.



а) запрос на определение авторов

б) запрос на проверку связи

Рисунок 4 – DL-запросы для проверки правильности добавления

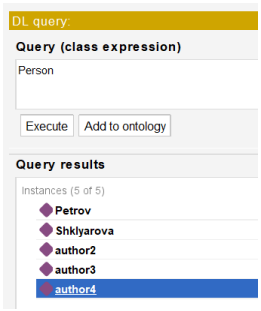


Рисунок 5 – DL-запрос, который выводит всех авторов

Таким образом, алгоритм позволяет расширить онтологию научно-учебных материалов, добавляя новые экземпляры и связи на основе данных из статей, обеспечивая корректность и согласованность с ожидаемой структурой и содержимым онтологии. Модуль для онтологии научно-учебных материалов позволит извлекать экземпляры классов, атрибуты экземпляров и информацию о связях между экземплярами, обогащая онтологию и давая более полное представление о предметной области.

Выводы

В статье была представлена методика формирования новых экземпляров в онтологиях, которая позволяет расширять и обогащать онтологии с учетом новых данных и связей. Процесс автоматического извлечения данных из научных статей и добавления их в онтологию с использованием семантических правил и шаблонов обеспечивает эффективное и надежное обновление онтологической модели. Проверка правильности результатов с помощью DL-запросов позволяет убедиться в корректности добавления экземпляров и установления связей. Эта методика имеет практическое применение для расширения онтологий и обеспечения точной и согласованной информации в информационных системах

Литература

1. Андриевская, Н. К. Онтологический подход в системах обработки данных научных и научно-образовательных организаций / Н. К. Андриевская // Проблемы искусственного интеллекта. – 2020. – № 1(16). – С. 23-36. – EDN WIZMRV.

2. Bird S., Klein E., Loper E. Natural Language Processing with Python // O'Reilly Media, 2009. – URL: https://www.researchgate.net/publication/220691633_Natural_Language_Processing_with_Python (дата посещения: 20.12.2022).

3. Официальная документация RDFLib. – URL: https://rdflib.readthedocs.io/ (дата посещения: 20.02.2023).

4. Официальная документация Owlready2. – URL: https://owlready2.readthedocs.io/ (дата посещения: 20.02.2023).

5. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M. Ontologies and Databases: The DL-Lite Approach // Cambridge University Press, 2017. – PP. 255-356.

6. Репозиторий PyOWL на GitHub. – URL: https://github.com/NeuralVedant/PyOWL (дата посещения: 20.02.2023).

7. Corcho, O., Fernández-López, M., Gómez-Pérez, A. Methodologies, tools and languages for building ontologies: Where is their meeting point? // Data & knowledge engineering, 2003. 46(1). – PP. 41-64. – URL: https://typeset.io/papers/methodologies-tools-and-languages-for-building-ontologies-3o724oh3up (дата посещения: 06.06.2023).

8. Шклярова, Е. Ю. Проектирование автоматизированной системы онлайн-поиска попутчиков / Е. Ю. Шклярова, С. Ю. Землянская, Е. Н. Мащенко // Информатика, управляющие системы, математическое и компьютерное моделирование (ИУСМКМ-2021) : Материалы XII Международной научно-технической конференции в рамках VII Международного Научного форума Донецкой Народной Республики к 100-летию ДонНТУ, Донецк, 26–27 мая 2021 года. – Донецк: Донецкий национальный технический университет, 2021. – С. 115-122. – EDN VVRAMZ.

9. Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E. Ontology Engineering in a Networked World // Ontology Engineering in a Networked World, 2012. – PP. 213-233.

10. Платонов А. В., Полещук Е. А. Методы автоматического построения онтологий. – URL: https://cyberleninka.ru/article/n/metody-avtomaticheskogo-postroeniya-ontologiy/viewer (дата обращения: 06.06.2023).

Шклярова Е. Ю., Землянская С. Ю. Формирование новых экземпляров в онтологии научной и учебно-методической информации. *Данная научная статья представляет исследование, посвященное разработке и применению онтологии научной и учебно-методической информации. В статье рассматриваются важные аспекты развития онтологии, представлена методика формирования новых экземпляров в существующей модели. Этот процесс позволяет расширить онтологию, улучшить процессы поиска и анализа информации. Представлен программный модуль, разработанный с использованием библиотеки RDFLib, который позволяет автоматически создавать новые экземпляры на основе данных, извлеченных из научных публикаций.*

Ключевые слова: *онтология, модификация онтологии, семантическая модель, новые экземпляры, концепт, сопоставление, целостность методика*

Shklyarova E., Zemlyanskaya S. Formation of New Instances in the Ontology of Scientific and Educational Methodological Information. *This research article presents a study dedicated to the development and application of an ontology for scientific and educational methodological information. The article examines important aspects of ontology development and introduces a methodology for forming new instances within an existing model. This process allows for the expansion of the ontology and improvement of information search and analysis processes. A software module, developed using the RDFLib library, is presented, which enables the automatic creation of new instances based on data extracted from scientific publications.*

Keywords: *ontology, ontology modification, semantic model, new instances, concepts, mapping, integrity, methodology.*

Статья поступила в редакцию 21.05.2023.
Рекомендована к публикации профессором Скобцовым Ю.А.

Об авторах

Аверин Геннадий Викторович – доктор технических наук, профессор, заведующий кафедрой компьютерных технологий ФГБОУ ВО «Донецкий государственный университет», г. Донецк.

Андриевская Наталия Климовна - кандидат технических наук, доцент кафедры автоматизированных систем управления факультета информационных систем и технологий ФГБОУ ВО «Донецкий национальный технический университет».

Вовченко Владислав Олегович - магистрант кафедры автоматизированных систем управления факультета информационных систем и технологий ФГБОУ ВО «Донецкий национальный технический университет».

Григорьев Александр Владимирович - кандидат технических наук, доцент, доцент кафедры программной инженерии им. Л. П. Фельдмана факультета интеллектуальных систем и программирования ФГБОУ ВО «Донецкий национальный технический университет».

Землянская Светлана Юрьевна - кандидат технических наук, доцент, доцент кафедры автоматизированных систем управления факультета информационных систем и технологий ФГБОУ ВО «Донецкий национальный технический университет».

Илюшин Павел Алексеевич, главный специалист, филиал АО «ЦЭНКИ» – «НИИ ПМ имени В.И. Кузнецова», г. Москва, Россия.

Куртенков Роман Владимирович - кандидат технических наук, доцент кафедры металлургии Санкт-Петербургского горного университета.

Личман Антон Александрович – аспирант кафедры компьютерной инженерии факультета интеллектуальных систем и программирования ФГБОУ ВО «Донецкий национальный технический университет».

Наумченко Владислав Павлович - инженер 1-й категории, филиал АО «ЦЭНКИ» – «НИИ ПМ имени В.И. Кузнецова», г. Москва, Россия.

Пикунов Дмитрий Григорьевич - начальник отдела, филиал АО «ЦЭНКИ» – «НИИ ПМ имени В.И. Кузнецова», г. Москва, Россия.

Приходченко Екатерина Ильинична – доктор педагогических наук, профессор, заведующий кафедрой инженерной педагогики и лингвистики ФГБОУ ВО «Донецкий национальный технический университет», заслуженный учитель Украины, академик Международной академии наук педагогического образования.

Руденко Мария Павловна - кандидат технических наук, доцент кафедры компьютерного моделирования и дизайна факультета информационных систем и технологий ФГБОУ ВО «Донецкий национальный технический университет».

Рычка Ольга Валентиновна – кандидат технических наук, доцент кафедры программной инженерии им. Л. П. Фельдмана факультета интеллектуальных систем и программирования ФГБОУ ВО «Донецкий национальный технический университет».

Сизякова Екатерина Викторовна - кандидат технических наук, доцент, доцент кафедры металлургии Санкт-Петербургского горного университета.

Светличная Виктория Антоновна - кандидат технических наук, доцент, доцент кафедры автоматизированных систем управления факультета информационных систем и технологий ФГБОУ ВО «Донецкий национальный технический университет».

Слободин Виктор Андреевич - студент, кафедра металлургии Санкт-Петербургского горного университета.

Соловьёв Алексей Владимирович - кандидат технических наук, начальник отделения, филиал АО «ЦЭНКИ» – «НИИ ПМ имени В.И. Кузнецова», г. Москва, Россия.

Толстых Маргарита Анатольевна – аспирант кафедры компьютерных технологий ФГБОУ ВО «Донецкий государственный университет», г. Донецк.

Филипишин Дмитрий Александрович – аспирант, ассистент кафедры программной инженерии имени Л.П. Фельдмана факультета интеллектуальных систем и программирования ФГБОУ ВО «Донецкий национальный технический университет».

Чередникова Ольга Юрьевна – кандидат технических наук, доцент, доцент кафедры компьютерной инженерии факультета интеллектуальных систем и программирования ФГБОУ ВО «Донецкий национальный технический университет».

Шклярова Екатерина Юрьевна - магистрант кафедры автоматизированных систем управления факультета информационных систем и технологий ФГБОУ ВО «Донецкий национальный технический университет».

**Требования к статьям,
направляемым в редакцию научного журнала
«Информатика и кибернетика»**

Редколлегией принимаются к рассмотрению статьи, в которых рассматриваются важные вопросы в области информатики и кибернетики. Научный журнал издаётся с 2015 года, периодичность издания – 4 раза в год.

В журнале предусмотрены следующие рубрики:

- информатика и вычислительная техника;
- компьютерные и информационные науки;
- инженерное образование.

В соответствии с номенклатурой специальностей научных работников МОН ДНР первые две рубрики соответствуют следующим укрупненным группам специальностей научных работников:

05.01 – «Инженерная геометрия и компьютерная графика»,

05.13 – «Информатика, вычислительная техника и управление».

С 01.02.2019 Научный журнал включён в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание учёной степени кандидата наук, на соискание учёной степени доктора наук (приказ МОН ДНР № 135) по группам специальностей 05.01.00 и 05.13.00.

Рубрика «Инженерное образование» предназначена опубликования сотрудниками научно-методических статей.

Журнал также включён в базу данных РИНЦ (Российский индекс научного цитирования) (лицензионный договор № 425-07/2016 от 14.07.2016).

Статьи, представляемые в данный сборник, должны отвечать следующим требованиям. ***Содержание статьи*** должно быть посвящено актуальным научным проблемам и включать следующие необходимые элементы:

- постановку проблемы в общем виде, её связь с важными научными и практическими задачами;
- анализ последних исследований и публикаций, в которых решается данная задача и на которые опирается автор, выделение нерешенных ранее частей общей проблемы, которым посвящается статья;
- формулировка цели статьи и постановка задач, решаемых в ней;
- изложение основного материала с полным обоснованием полученных научных результатов;
- выводы и перспективы последующих исследований в данном направлении.

Каждый элемент должен быть выделен соответствующим названием раздела, например, «введение», «постановка задачи», «цель и задачи работы», «цель статьи», «цель исследования», «цель разработки», «анализ … », «сравнительная оценка … », «разработка … », «проектирование … », «программная реализация», «тестирование … », «полученные результаты», «выводы», «литература». Разделы «введение», «выводы», «литература» являются обязательными. Включать в названия разделов нумерацию не разрешается.

В основном тексте статьи формулируются и обосновываются полученные авторами утверждения и результаты. Выводы должны полностью соответствовать содержанию основного текста. Языки публикаций: русский, английский.

Объём статьи, формат страницы

Для оформления статьи следует использовать листы формата А4 (210х297 мм) с полями по 2,5 см со всех сторон. Нумерацию страниц выполнять не нужно.

Рекомендуемый объём статьи – 6-12 страниц. Рукописи меньшего объёма могут быть рекомендованы к публикации в качестве коротких сообщений.

Последняя страница текста статьи должна быть заполнена не менее чем на две трети, но содержать не менее трёх пустых строк в конце.

Форматирование текста

Подготовка статьи осуществляется в текстовом редакторе Microsoft Office Word.

Весь текст статьи оформляется шрифтом Times New Roman 10 пт с одинарным междустрочным интервалом, если ниже в требованиях не сказано иного. Абзацный интервал «перед» – 0 пт, «после» – 0 пт.

На первой строке с выравниванием по левому краю располагается УДК.

Заголовок (название) статьи оформляется шрифтом Times New Roman 14 пт, полужирное начертание, с выравниванием по центру (без абзацных отступов). Заголовок статьи следует печатать с прописной буквы без точки в конце, переносы слов не допускаются. Абзацный интервал «перед» – 12 пт, «после» – 12 пт.

После названия статьи следует информация об авторах, которая выравнивается по центру (без абзацных отступов). На одной строке указываются инициалы и фамилии всех авторов через запятую. Между двумя инициалами ставится пробел. С новой строки указывается название вуза (организации) и город (для каждого автора, если не совпадают). На следующей строке указываются адреса электронной почты (один адрес либо каждого автора – по желанию). Адрес электронной почты оформляется в виде гиперссылки.

К тексту аннотации применяется курсивное начертание, с выравниванием по ширине, отступы слева и справа по 1 см. Заголовок «Аннотация» выделяется полужирным начертанием. Объём аннотации – 450-550 символов (без пробелов). Абзацный интервал «перед» – 12 пт, «после» – 12 пт.

Основной текст статьи разбивается на две колонки шириной по 7,5 см (промежуток между столбцами – 0,99 см), выравнивается по ширине. Абзацный отступ первой строки – 1 см. Автоматический перенос слов не применяется.

Заголовки разделов выполняются шрифтом Arial 10 пт, полужирное курсивное начертание. Абзацный отступ отсутствует, интервал перед абзацем – 12 пт, после абзаца – 6 пт. Для заголовка «Введение» установить интервал «перед» – 0 пт, «после» – 6 пт.

Таблицы в тексте статьи

Название следует помещать над таблицей с абзацного отступа (1 см) в формате: слово «Таблица», пробел, номер таблицы, пробел, тире, пробел, название таблицы. Название таблицы записывают с прописной буквы без точки в конце строки и выравнивают по ширине. В ячейках таблицы устанавливается выравнивание текста по центру по вертикали. По горизонтали текст выравнивается по центру либо по левому краю. Границы ячеек таблицы должны быть только чёрного цвета, толщина линии – 1 пт. На все таблицы должны быть приведены ссылки в тексте статьи, при ссылке следует писать слово «табл.» с указанием её номера, например, « … данные приведены в табл. 5». Таблицы нумеруются в пределах статьи. Таблица располагается сразу после ссылки на неё, если это возможно (например, после окончания абзаца). Если же таблица не помещается на текущей странице, то она должна быть расположена в начале следующей страницы (или колонки). При необходимости допускается включение в статью таблицы, ширина которой превышает ширину колонки. В этом случае таблица и её название размещаются по центру страницы. Таблица не должна выступать за границы полей страницы. Таблица и её название отделяются от основного текста статьи одной пустой строкой до и после.

Рисунки в статье

Ссылки на иллюстрации по тексту статьи обязательны и оформляются в виде « ... на рис. 2» и т. п. Рисунок и его подпись выравниваются по центру колонки (без абзацных отступов), положение рисунка – «в тексте». Размещается рисунок после его первого упоминания в тексте, если это возможно (например, после окончания абзаца). Если же иллюстрация не помещается на текущей странице, то она должна быть расположена в начале следующей страницы (или колонки). При необходимости допускается включение в статью рисунка, ширина которого превышает ширину колонки. В этом случае рисунок и его подпись выравниваются по центру страницы. Иллюстрация не должна выступать за границы полей страницы. Подпись рисунка оформляется в формате: слово «Рисунок», пробел, номер иллюстрации, пробел, тире, пробел, название рисунка. Название рисунка записывают с прописной буквы без точки в конце строки. Для подписи иллюстрации применяют курсивное

начертание. Иллюстрация и её подпись отделяются от основного текста статьи одной пустой строкой до и после. Не допускается выполнять рисунки с помощью встроенного графического редактора Microsoft Office Word. Если на иллюстрации имеется текст, размер шрифта должен быть не менее чем аналогичный текст, набранный шрифтом Times New Roman 10-го размера. Иллюстрация не должна содержать много незаполненного пространства.

Формулы

Формулы и уравнения рекомендуется набирать с использованием MathType (предпочтительно) или MS Equation. Формулы и математические символы не должны существенно отличаться по размеру от основного текста. Обязательной является нумерация формул, на которые имеется ссылка в тексте статьи. Ссылки в тексте на порядковые номера формул дают в скобках, например, « ... согласно формуле (2)». Формулы размещаются по центру колонки, а их номера – по правому краю. Как для строки с формулой, так и для первой строки пояснений (при наличии), абзацный отступ убирается. Первая строка пояснения начинается со слова «где», после которого следует поставить табуляцию на 1 см, затем само пояснение в формате: символ, подлежащий объяснению, пробел, тире, пробел, поясняющий текст, запятая, обозначение единицы измерения физической величины. Пояснения перечисляются через точку с запятой, выравниваются по ширине. Вторая и последующие строки пояснений начинаются с абзацного отступа (1 см). Весь блок текста, связанный с формулой (только формула, несколько формул подряд или формула с пояснениями), отделяется от основного текста одной пустой строкой до и после. Переносить формулы на следующую строку допускается только на знаках выполняемых операций, причем знак в начале следующей строки повторяют. При переносе формулы на знаке умножения применяют знак «×». Формулы и математические уравнений могут быть записаны в тексте документа, если их высота не превышает высоту строки. При этом следует учитывать, что знаки математических операций отделяются от чисел или символов пробелами с обеих сторон. Например, «Если учесть, что $y < 0$ и $2x + y = 1$, то из формулы (3) можно выразить x…». К символам, которые приведены в формуле, при дальнейшем их употреблении (в том числе в пояснениях к формуле) должно применяться курсивное начертание. При этом к любым числам (верхние и нижние индексы, содержащие цифры и т. п.), а также к математическим знакам курсивное начертание не применяется. Не допускается вставлять формулы, выполненные в виде рисунков.

Перечисления: оформление списков

Основной текст статьи может содержать перечисления, оформленные в виде маркированного списка. В качестве маркера элемента списка разрешается использовать только короткое тире «–». Каждый элемент перечисления записывается с новой строки с абзацного отступа, равного 1 см. После символа короткого тире текст располагается с отступом в 1,5 см от левой границы строки, выравнивается по ширине, при переносе на новые строки располагается без отступов. Нумерованные и многоуровневые списки включать в статью не разрешается.

Литература

В тексте статьи обязательны ссылки на все литературные источники, номер источника указывается в квадратных скобках. Ссылки на неопубликованные работы не допускаются. Рекомендуемое количество источников, на которые ссылается автор, не менее 10. Перечень источников приводится в порядке их упоминания в статье. Библиографическое описание каждого литературного источника оформляется в соответствии с ГОСТ Р 7.0.100–2018. Перечень литературных источников оформляется в виде нумерованного списка. В качестве маркеров элементов списка используют порядковые арабские цифры с точкой. Каждый источник представляет собой отдельный элемент перечисления, записывается с новой строки с абзацного отступа, равного 1 см. После порядкового номера с точкой текст располагается с отступом в 1,5 см от левой границы строки, выравнивается по ширине, при переносе на новые строки располагается без отступов.

В конце статьи обязательно приводятся аннотации на русском и английском языках, каждая заканчивается перечнем 5-6 ключевых слов.

К тексту аннотации применяется курсивное начертание, с выравниванием по ширине, отступы слева и справа по 1 см. Слово «Аннотация» опускается. Текст аннотации начинается с ФИО авторов и названия статьи, выделяемых полужирным начертанием. Аннотация на русском языке совпадает с аннотацией, приведенной в начале статьи. В тексте аннотации на английском языке после фамилии автора указывается только первая буква имени с точкой. Абзацный интервал «перед» – 12 пт, «после» – 12 пт. Ключевые слова оформляются с новой строки аналогично тексту аннотации. Заголовок «Ключевые слова:» (англ. «Keywords:») выделяется полужирным начертанием. Ключевые слова перечисляются через запятую.

Порядок представления статьи и сопроводительные документы

В редакцию необходимо представить:

– файл с текстом статьи;

– файл, содержащий фамилию, имя и отчество авторов полностью; ученую степень, ученое звание; место работы с полным указанием должности, подразделения и наименования организации, города (страны); номера телефонов и e-mail для связи;

– экспертное заключение о возможности публикации статьи, подписанное руководителем и заверенное печатью организации, в которой работает автор статьи;

– выписка из заседания кафедры или письмо организации с просьбой об опубликовании и указанием, что изложенные в статье результаты ранее не публиковались.

Статьи и сопроводительные документы следует высылать на электронный адрес infcyb.donntu@yandex.ru.

К сведению авторов

Если статья оформлена с нарушением указанных выше требований и правил, редакция после предварительного рассмотрения может отклонить статью.

На рецензирование статьи направляются членам редакционной коллегии журнала. Все статьи публикуются при наличии положительной рецензии.

В статью могут быть внесены изменения редакционного характера без согласования с автором. Ответственность за содержание статьи и качество перевода аннотаций несут авторы.

Публикация статей в научном журнале «Информатика и кибернетика» осуществляется на некоммерческой основе.

Все номера Научного журнала размещаются на сайте http://infcyb.donntu.ru/.

.

CONTENT

Informatics and computer engineering

Электронное периодическое издание

Научный журнал

ИНФОРМАТИКА И КИБЕРНЕТИКА

(на русском, английском языках)

№ 2 (32) - 2023

Ответственный за выпуск Р. В. Мальчева

Технический редактор Р. В. Мальчева

Компьютерная верстка Р. В. Мальчева

Подписано к выпуску 18.06.2023. Усл. печ. лист. 7,5. Уч.-изд. лист.4,6.
Адрес редакции: ДНР, 283001, г. Донецк, ул. Артема, 58, ФГБОУ ВО «ДонНТУ»,
4-й учебный корпус, к. 36, ул. Кобозева, 17.
Тел.: +7 (856) 301-07-35, +7 (949) 334-89-11
E-mail: infcyb.donntu@yandex.ru, URL: http://infcyb.donntu.ru