



Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____ «Информатика и системы управления»

КАФЕДРА _____ «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

НА ТЕМУ:

«Метод построения поисковых индексов в реляционной
базе данных на основе глубоких нейронных сетей»

| | | | |
|-----------------|----------------------------|--------------------------|---|
| Студент: | <u>ИУ7-83Б</u> (группа) | _____ (подпись, дата) | <u>М. Д. Маслова</u> (И. О. Фамилия) |
| Руководитель: | | _____ (подпись, дата) | <u>А. А. Оленев</u> (И. О. Фамилия) |
| Нормоконтролер: | | _____ (подпись, дата) | _____ (И. О. Фамилия) |

2023 г.

СОДЕРЖАНИЕ

| | | |
|----------|--|-----------|
| 1 | Конструкторская часть | 4 |
| 1.1 | Требования и ограничения метода | 4 |
| 1.2 | Особенности метода построения индекса | 4 |
| 1.2.1 | Общее описание метода построения индекса | 4 |
| 1.2.2 | Предварительная обработка данных | 6 |
| 1.2.3 | Разработка архитектуры глубокой нейронной сети | 8 |
| 1.3 | Разработка алгоритмов поиска и вставки | 11 |
| 1.4 | Данные для обучения и тестирования индекса | 12 |
| 2 | Технологическая часть | 14 |
| 2.1 | Выбор средств программной реализации | 14 |
| 2.2 | Реализация программного обеспечения | 14 |
| 2.2.1 | Форматы входных и выходных данных | 14 |
| 2.2.2 | Поддерживаемые виды запросов | 15 |
| 2.2.3 | Программный интерфейс виртуальных таблиц | 15 |
| 2.2.4 | Реализация индекса | 15 |
| 2.3 | Сборка программного обеспечения | 15 |
| 2.4 | Взаимодействие с программным обеспечением | 15 |
| 2.5 | Пример работы | 15 |
| 2.6 | Результаты тестирования | 15 |

1 Конструкторская часть

1.1 Требования и ограничения метода

Метод построения поисковых индексов в реляционной базе данных на основе глубоких нейронных сетей (далее – метод построения индексов) должен:

1. получать из таблицы реляционной базы данных набор ключей и набор соответствующих указателей на записи в индексируемой таблице реляционной базы данных или иных значений, выполняющих роль указателей;
2. выполнять предварительную обработку полученных наборов, такую, как их совместную сортировку по значениям ключей, получение позиций ключей в отсортированном виде и нормализацию ключей и позиций;
3. обучать модель нейронной сети на подготовленном наборе ключей и позиций;
4. сохранять параметры обученной модели для каждой таблицы с целью возможности выполнять запросы поиска без переобучения;
5. обеспечивать поиск записи (диапазона записей) таблицы по ключу (диапазону ключей) с использованием обученной модели;
6. обеспечивать корректность операции поиска после вставки/удаления новых записей путем переобучения модели;

На разрабатываемый метод накладываются следующие ограничения:

- в качестве ключей на вход принимаются целые числа для исключения решения дополнительной задачи преобразования входных данных;
- ключи во входном наборе уникальны.

1.2 Особенности метода построения индекса

1.2.1 Общее описание метода построения индекса

Основные этапы метода построения индекса приведены на функциональной декомпозиции метода на рисунке 1.1.

На вход методу подается набор уникальных целочисленных ключей, которые перед обучением модели глубокой нейронной сети проходят предварительную обработку по определенным правилам, описанным далее. Отдельным

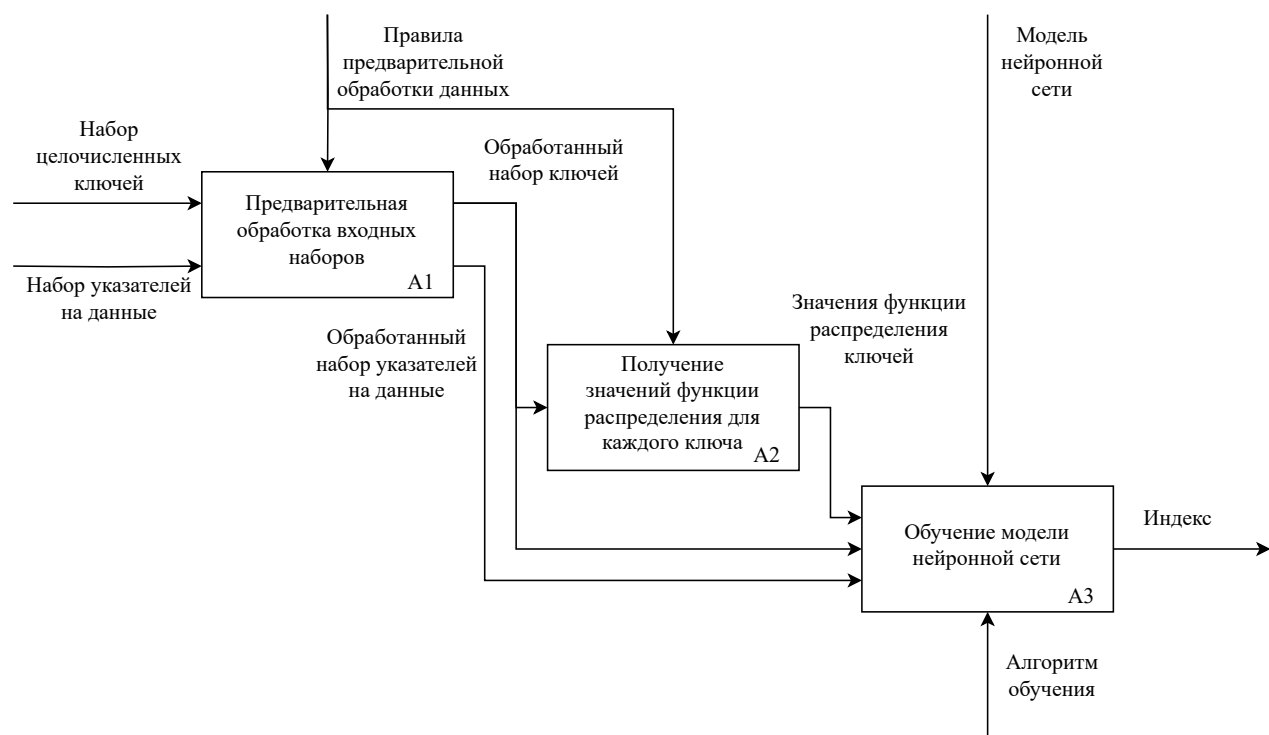


Рисунок 1.1 – Функциональная схема метода построения индекса

этапом выделено получение значений функций распределения для каждого ключа, относящееся к предварительной обработке, но представляющее собой ее ключевой момент. Полученные после первых двух этапов обработанные ключи и соответствующие значения функций используются для обучения модели глубокой нейронной сети в качестве признаков и меток соответственно.

Ключевым моментом метода является представление в отсортированном (по ключам) виде наборов ключей и набора соответствующих указателей на данные. Именно отсортированный вид позволяет использовать закономерность распределения ключей по позициям для обучения модели, предсказывать позиции ключей и уточнять их.

Результатом работы метода является структура данных, представляющая собой индекс на основе глубокой нейронной сети и имеющая следующие поля:

- отсортированный массив ключей, поданных на вход;
- отсортированный по значениям ключей массив указателей на данные, соответствующие ключам;
- модель обученной глубокой нейронной сети, с помощью которой будет предсказываться положение ключа в отсортированном массиве;
- средняя и максимальная абсолютные ошибки предсказания позиции

ключа, для ее уточнения и возврата верного указателя на данные.

Краткое описание индекса, являющегося результатом работы метода, как структуры данных представлено на рисунке 1.2.

| Индекс | |
|------------|-------------------------|
| - model | : модель нейронной сети |
| - keys | : массив целых чисел |
| - data | : массив указателей |
| - max_err | : целое число |
| - mean_err | : целое число |

Рисунок 1.2 – Индекс как структура данных

Подробное описание каждого этапа приведено в следующих пунктах данного подраздела.

1.2.2 Предварительная обработка данных

Разрабатываемый метод построения индекса предполагает предварительную обработку набора целочисленных ключей, схема алгоритма которой представлена на рисунке 1.3.

На вход подаются согласованные массивы ключей и указателей на данные, то есть считается, что ключ, стоящий на первой позиции в массиве ключей, идентифицирует данные по указателю, стоящему на первой позиции в массиве указателей; ключ, стоящий на второй, — указатель, стоящий на второй, и так далее. С учетом этого происходит согласованная сортировка двух массивов по значениям ключей.

Далее для последующего обучения модели глубокой нейронной сети производится нормализация ключей, выступающих в качестве входных данных сети, в диапазон $[0, 1]$, для чего используется метод минимакс-нормализации, при котором нормализованное значение вычисляется по формуле:

$$x_{\text{норм}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \quad (1.1)$$

где $x_{\text{норм}}$ — нормализованное значение ключа;

x — натуральное значение ключа;

x_{\min} , x_{\max} — минимальное и максимальное возможное значение ключа в

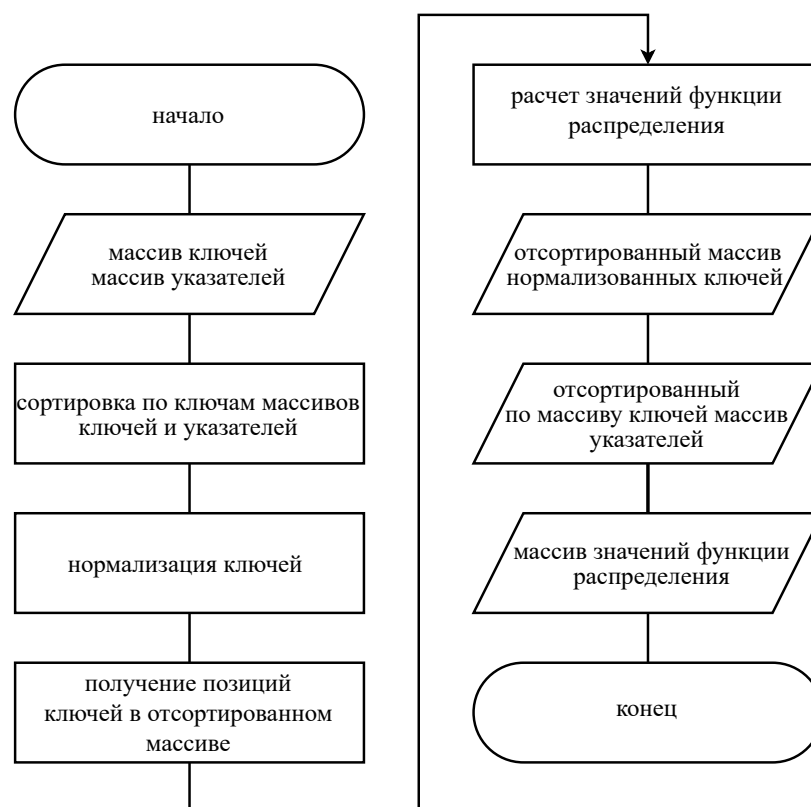


Рисунок 1.3 – Схема алгоритма предварительной обработки данных

наборе соответственно.

Далее полученный набор ключей размечается путем вычисления для каждого ключа K значения функции распределения F по его позиции P в отсортированном массиве и количества индексируемых ключей N с помощью формулы:

$$F(K) = \frac{P}{N} \quad (1.2)$$

На этом предварительная обработка завершается и полученные отсортированные массивы ключей и указателей, а также соответствующие значения функции распределения передаются в качестве входных данных на этап обучения модели глубокой нейронной сети.

Полное описание алгоритма предварительной обработки представлено на листинге 1.1.

Листинг 1.1 – Предварительная обработка данных

Вход:

keys : массив целочисленных ключей;
data : массив указателей на данные, соответствующие ключам;
N : длина массивов.

Выход:

keys : отсортированный массив нормализованных ключей;
data : отсортированный по массиву ключей массив указателей;
cdf : массив значений функции распределения.

```
1 begin
2   сортировать keys и data по keys;
   ▷ здесь и далее под операцией к вектору (массиву) и числу понимается
   ▷ применение данной операции с данным числом к каждому элементу вектора
3    $keys \leftarrow \frac{keys - keys[0]}{keys[N-1] - keys[0]}$ ;
4    $positions \leftarrow [0, 1, \dots, N - 1]$ ;
5    $cdf \leftarrow \frac{positions}{N-1}$ ;
6   return keys, data, cdf;
7 end
```

1.2.3 Разработка архитектуры глубокой нейронной сети

Полученные на этапе предварительной обработки массивы ключей и соответствующих им значений функций распределения используются для обучения глубокой нейронной сети. Массив данных, хранимый в индексе, используется для возврата нужных данных при поиске, но не для обучения.

Задача построения индекса на основе нейронной сети сводится к задаче аппроксимации функции распределения, для решения которой подходят полносвязные нейронные сети.

За основу архитектуры глубокой нейронной сети принята архитектура из исследования [1D2D], которая представлена на рисунке 1.4. Это полносвязная нейронная сеть с двумя скрытыми слоями по 32 нейрона. В качестве функции активации в каждом нейроне скрытых слоев используется линейный выпрямитель или ReLU (*Rectified Linear Unit*), значение которой вычисляется по формуле 1.3, чему соответствует график, представленный на рисунке 1.5.

$$f(x) = \max(0, x). \quad (1.3)$$

Активационная функция выходного слоя является линейной.

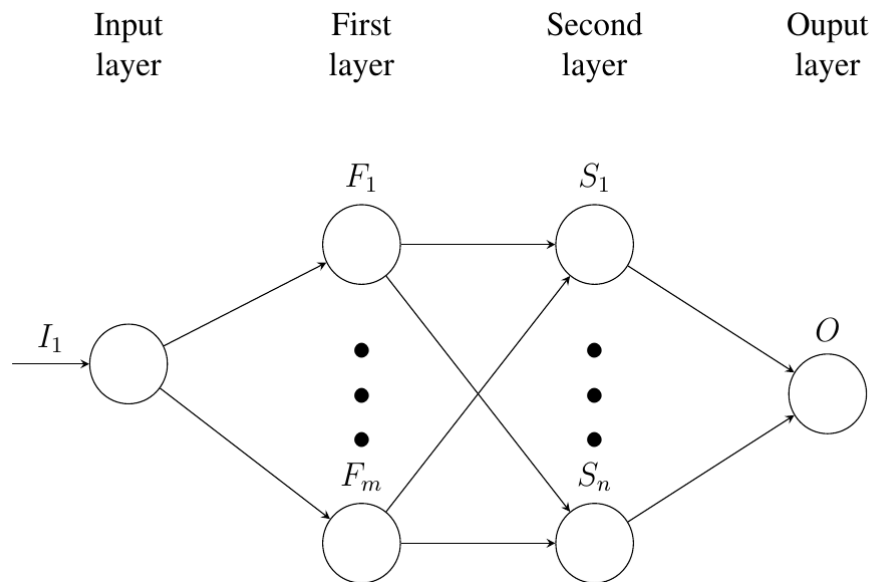


Рисунок 1.4 – Полносвязная нейронная сеть с двумя скрытыми слоями

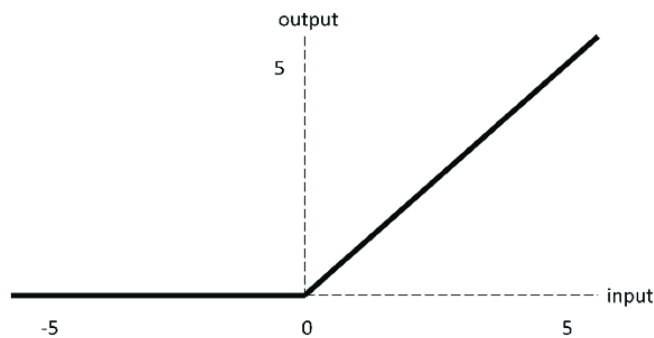


Рисунок 1.5 – График функции ReLU

Для исследования возможности увеличения точности предсказания, и как следствие уменьшение времени поиска, в качестве модели глубокой нейронной сети, представляющей основу индекса, используется полносвязная нейронная сеть с тремя слоями, представленная на рисунке 1.6. Число нейронов в слоях и активационные функции приняты такими же, как в случае глубокой нейронной сети с двумя скрытыми слоями.

Обучение обеих моделей глубокой нейронной сети начинается с инициализации весов случайно сгенерированными значениями по распределению $U(-\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}})$. Собственно обучение проводится методом стохастического градиентного спуска с оптимизацией в качестве функции потерь среднеквадратической ошибки (MSE — *mean squared error*). Описание алгоритма обучения приведено на листинге 1.2.

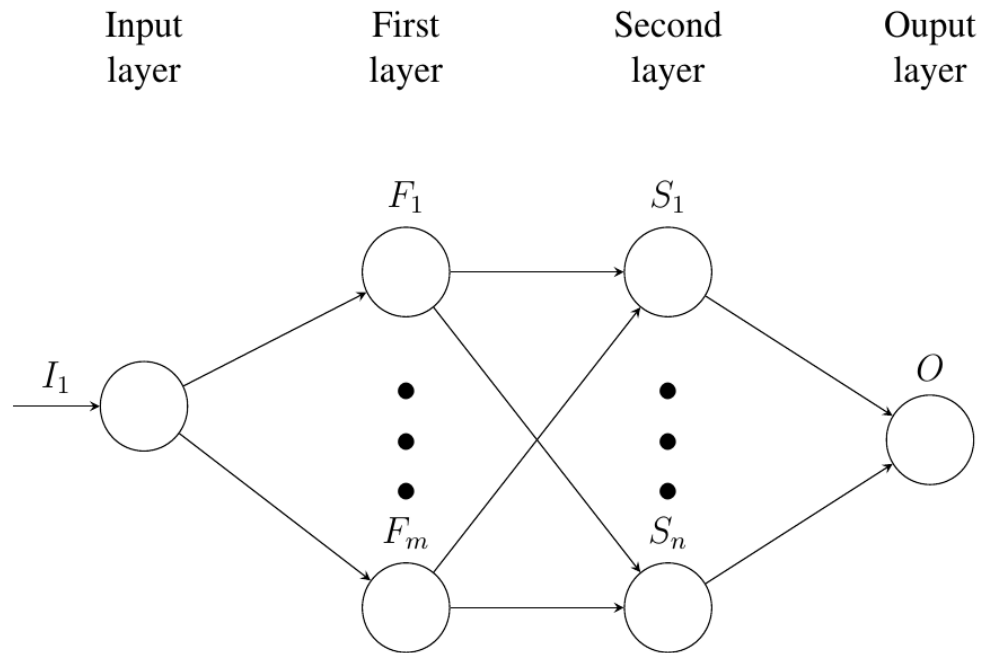


Рисунок 1.6 – Полносвязная нейронная сеть с тремя скрытыми слоями

Листинг 1.2 – Алгоритм обучения глубокой нейронной сети на основе градиентного спуска

Вход:

keys : массив нормализованных ключей;
cdf : массив значений функции распределения для каждого ключа;
N : длина массивов;
epochs : количество эпох;
alpha : скорость обучения.

Выход:

model : обученная модель.

```

1 begin
2   model  $\leftarrow$  структура модели;
    $\triangleright$  вектор смещений ключей в матрицу весов
3   model.weights  $\leftarrow U(-\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}})$  ;  $\triangleright$  случайные значения из распределения
4   for epoch  $\leftarrow 1$  to epochs do
5     согласованно перемешать keys и cdf;
6     for i  $\leftarrow 0$  to N - 1 do
7        $\hat{y} = model.predict(key[0])$  ;  $\triangleright$  прямой проход
8       mse =  $(\hat{y} - cdf[i])^2$ ;
9       gradients  $\leftarrow backward\_pass(model, mse)$  ;  $\triangleright$  обратный проход
10      model.weights  $\leftarrow model.weights - alpha \cdot gradients$ 
11    end
12  end
13 end

```

Так как в случае поискового индекса глубокая нейронная сеть будет предсказывать положения только тех ключей, на которых она обучалась, явления переобучения нейронной сети является положительным, поэтому в качестве одного батча выступает одна пара (ключ; значение функции распределения), что также отражено на листинге выше.

1.3 Разработка алгоритмов поиска и вставки

Основной операцией, выполняемой с помощью индекса, является поиск, функциональная схема выполнения которого представлена на рисунках 1.7-1.8.

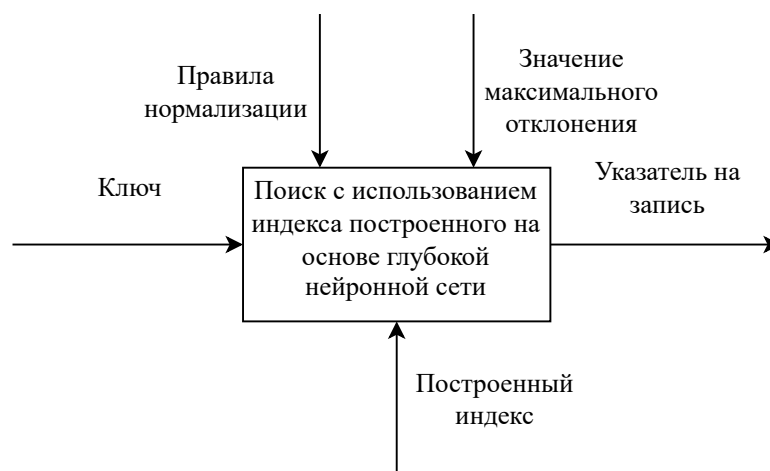


Рисунок 1.7 – Функциональная схема нулевого уровня поиска

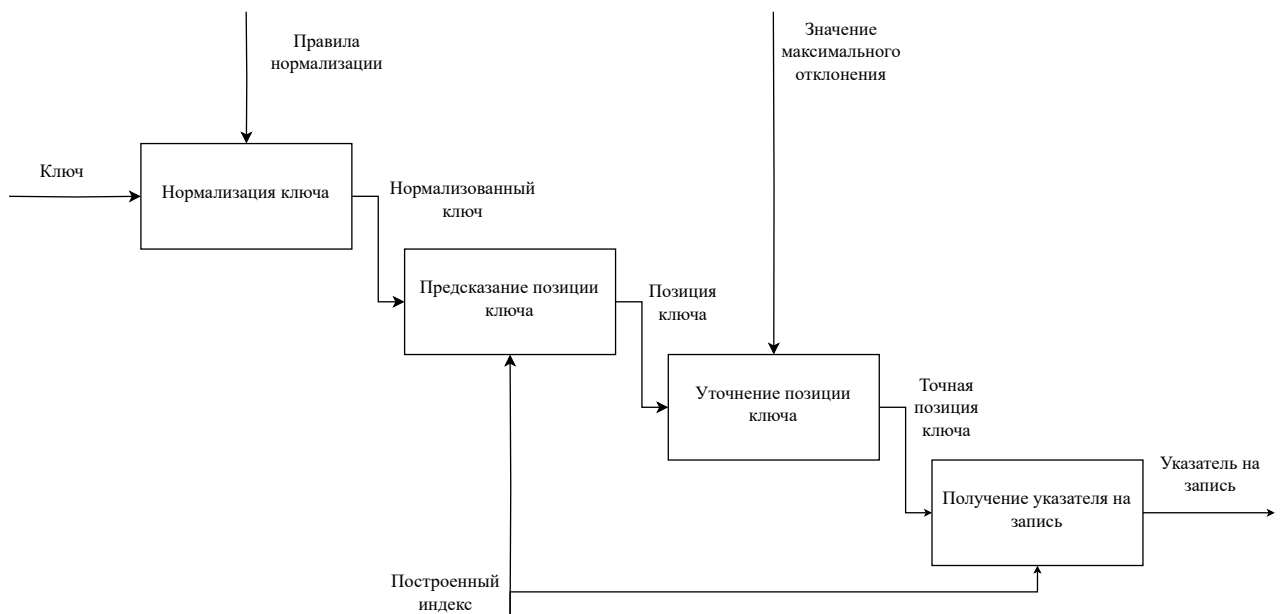


Рисунок 1.8 – Функциональная схема первого уровня поиска

Получаемая позиция требует уточнения, происходящего за счет получаемого в результате обучения модели максимального отклонения от истинного расположения. (Алгоритм уточнения???)

Для реализации вставки происходит добавление новых значений ключа и указателя в существующие массивы, и повторяется алгоритм построения индекса (??? не сначала, а со значений параметров уже обученной модели ???).

1.4 Данные для обучения и тестирования индекса

Так как в основе индекса на основе глубоких нейронных сетей лежит аппроксимация функции распределения ключей, работу разработанного метода необходимо протестировать на различных законах, перечисленных далее.

- Равномерный закон $R[a, b]$, функция распределения которого описывается формулой 1.4.

$$F(x) = \begin{cases} 0, & \text{если } x < a \\ \frac{x-a}{b-a}, & \text{если } a \leq x \leq b \\ 1, & \text{если } x > b. \end{cases} \quad (1.4)$$

Нормализованные ключей лежат в диапазоне $[0, 1]$, значения функции распределения за пределами этого диапазона не представляют интереса для построения индекса, поэтому можно считать, что функция имеет вид, представленный формулой 1.5.

$$F(x) = x, \quad x \in [0, 1]. \quad (1.5)$$

- Нормальный закон $N(\mu, \sigma^2)$, функция распределения которого описывается формулой 1.6.

$$F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt \quad (1.6)$$

Для тестирования метода построения по заданным законам генерируются значения ключей, по совокупности которых формируется эмпирическая функция распределения, как это было описано выше.

Также для проверки работы метода требуется оценить его работоспособность на реальных данных, в качестве которых выбраны уникальные идентифи-

каторы элементов из открытого набора географических данных *OpenStreetMap*, или *OSM* [**osm**], функция распределения которых имеет более сложный вид, чем основные законы распределения.

Графики функций распределения каждого из набора входных ключей, представлены на рисунке ??.

2 Технологическая часть

2.1 Выбор средств программной реализации

Для реализации метода построения индекса в реляционной базе данных на основе глубоких нейронных сетей в качестве языка программирования выбран Python 3.10 [**python**], так как предоставляет широкий выбор библиотек для глубокого обучения и визуализации его результатов. В качестве библиотеки глубокого обучения выбран TensorFlow 2.11.0 [**tf**] и работающий поверх нее высокоуровневый программный интерфейс Keras 2.11.0 [**keras**]. Для работы с массивами данных выбрана библиотека numpy [**numpy**].

В качестве реляционной системы управления базами данных выбрана SQLite [**sqlite**], предоставляющая программный интерфейс виртуальных таблиц, позволяющих релизовать пользовательский поисковый индекс. Виртуальные таблицы являются одним из видов расширений SQLite, программный интерфейс которых предоставляется на языке C [**c**], который и выбран в качестве языка программирования для взаимодействия с реляционной базой данных.

Для обеспечения взаимодействия между компонентом работы с базой данных и компонентом, непосредственно реализующий индекс, используются библиотеки языка C `Python.h` для работы с объектами языка Python и `numpy/arrayobject.h`, предоставляющая программный интерфейс для работы с numpy-массивами, которые являются основным типом данных, через который происходит взаимодействие модулей.

2.2 Реализация программного обеспечения

2.2.1 Форматы входных и выходных данных

Создание таблицы, для заполнения csv формат

Для создания индекса `id` столбца, значения которого должны быть представлены в виде целых чисел, строковое имя модели индекса, на основе которой он строится.

2.2.2 Поддерживаемые виды запросов

2.2.3 Программный интерфейс виртуальных таблиц

структура модуля

функция инициализации

создание и удаление, подключение и отключение таблицы

функции курсора `numru` массив

реализации

2.2.4 Реализация индекса

2.3 Сборка программного обеспечения

2.4 Взаимодействие с программным обеспечением

2.5 Пример работы

2.6 Результаты тестирования