

ності навчальних даних. In-Context Learning доцільно застосовувати в ситуаціях, коли є необхідність у швидкій адаптації без додаткового навчання [3]. Fine-tuning може забезпечити більш високий рівень точності, особливо коли є достатній обсяг даних для налаштування моделі. Meta-learning, попри свої ресурсоємність та складність реалізації, забезпечує найкращі результати в умовах, коли модель повинна часто адаптуватися до нових завдань з мінімальними витратами на навчання.

Проведений аналіз дає змогу зробити висновок, що техніки few-shot learning є ефективним інструментом для доменної адаптації великих мовних моделей в умовах обмежених даних. Вибір конкретної техніки залежить від конкретних потреб і обмежень. У тих випадках, коли важлива швидка адаптація, рекомендується використання In-Context Learning, тоді як для досягнення високої точності доцільним є застосування fine-tuning. У випадках з частою зміною доменів і потребою в адаптації на основі мінімальних даних найкращим вибором є meta-learning.

Список використаних джерел

1. Language Models are Few-Shot Learners / Т. В. Brown, В. Mann, N. Ryder Subbiah et al. URL: <https://arxiv.org/abs/2005.14165> (дата звернення: 07.11.2024).
2. Evaluating Large Language Models Trained on Code / М. Chen, J. Tworek, Н. Jun, Q. Yuan, P. de Oliveira et al. URL: <https://arxiv.org/abs/2107.03374> (дата звернення: 07.11.2024).
3. Finn C., Abbeel P., Levine S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *Proceedings of the 34th International Conference on Machine Learning*. URL: <https://arxiv.org/abs/1703.03400> (дата звернення: 07.11.2024).
4. Language Models are Unsupervised Multitask Learners / А. Radford., J. Wu., R. Child, D. Luan, D. Amodei, I. Sutskever. *OpenAI Blog*. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (дата звернення: 07.11.2024).
5. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale / А. Dosovitskiy, L. Beyer, А. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. URL: <https://arxiv.org/abs/2010.11929> (дата звернення: 07.11.2024).
6. Pre-trained Models for Natural Language Processing: A Survey / X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang. *Science China Technological Sciences*. 2021. № 63. P. 1872–1897. URL: <https://arxiv.org/abs/2003.08271> (дата звернення: 07.11.2024).

УДК 519.872.5:519.21(043.2)

*Чемес В. С., здобувач вищої освіти,
Сеник І. О., асистент кафедри
інформаційних технологій*

АНАЛІЗ І ЗАСТОСУВАННЯ МОДЕЛЕЙ ТЕОРІЇ ЧЕРГ У СИСТЕМАХ ОБСЛУГОВУВАННЯ

Донецький національний університет імені Василя Стуса, м. Вінниця

У сучасних умовах розвитку економіки та сектору послуг проблема ефективно організації систем обслуговування стає особливо важливою. Майже

щодня більшість людей стикаються з довгими чергами в магазинах, банках, лікарнях тощо. Неефективне обслуговування призводить до значних часових втрат клієнтів, зниження якості сервісу та економічних збитків для компанії.

Теорія черг – це математичний підхід до аналізу та оптимізації процесів обслуговування, коли потік клієнтів або об'єктів чекає на обслуговування на певних серверах або ресурсах. Для моделювання та дослідження різноманітних реальних ситуацій, як-от обслуговування клієнтів у банках, аеропортах, сервісних центрах або мережах передачі даних широко використовується теорія черг. За допомогою неї можна оптимізувати обслуговування клієнтів і витрати на утримання системи [1].

Основою теорії черг є декілька ключових понять, що описують поведінку клієнтів і системи загалом. Клієнт – це одиниця, яка потребує обслуговування і є потенційним учасником черги (наприклад, покупець у магазині або пакет даних у комп'ютерній мережі). Обслуговування надається спеціальними серверами, які, власне, і складають систему обслуговування, наприклад, касир у супермаркеті або сервер у дата-центрі. Потік клієнтів може бути випадковим або регулярним і визначається інтенсивністю, тобто кількістю клієнтів, які надходять у систему за одиницю часу (λ). Інтенсивність обслуговування відображає, скількох клієнтів сервер здатен обслужити за той самий проміжок часу (μ) [1, 2].

Системи черг класифікуються за типом потоку клієнтів, числом серверів і розподілом часу обслуговування. Наприклад, у позначенні M/M/1 кожна літера або цифра описує одну з характеристик системи.

Перша літера «M» (Markovian) означає, що потік клієнтів є випадковим і має розподіл Пуассона, що характеризує ситуації, коли клієнти надходять із певною середньою інтенсивністю, але з випадковими інтервалами між собою. Друга «M» вказує, що час обслуговування є експоненціально розподіленим, що характерно для процесів, у яких середня тривалість обслуговування постійна, але кожен окремий випадок може значно відхилитися. Остання цифра вказує кількість серверів: у моделі M/M/1 – один сервер, який обслуговує клієнтів поспідовно, тому черги можуть бути довшими за високої інтенсивності потоку.

У моделі M/M/s структура подібна, проте «s» позначає наявність кількох серверів, які працюють одночасно. Це дає змогу обслуговувати кількох клієнтів паралельно, знижуючи час очікування і підвищуючи пропускну здатність системи. Для моделі M/M/s, де є кілька серверів, основні формули для розрахунку характеристик системи змінюються і враховують кількість серверів [3, 4].

Середня кількість клієнтів у системі відображає середнє число клієнтів, які одночасно перебувають у системі, включно з тими, які чекають у черзі та обслуговуються:

$$L = \frac{\lambda}{\mu} \cdot \left[\frac{1}{(1 - p)} \right],$$

де коефіцієнт завантаження кожного сервера:

$$p = \frac{\lambda}{\mu \cdot s}.$$

Середній час перебування клієнта в системі вказує на час, який клієнт проводить у системі від моменту прибуття до завершення обслуговування:

$$W = \frac{1}{\mu} \cdot \left[\frac{1}{(1-p)} \right].$$

Середня кількість клієнтів у черзі – це кількість клієнтів, які очікують на обслуговування:

$$L_q = \frac{\lambda^2}{\mu \cdot (s \cdot (\mu - \lambda))} \cdot \left[\frac{1}{(1-p)} \right].$$

Середній час очікування в черзі відображає час, який клієнт проводить у черзі перед обслуговуванням:

$$W_q = \frac{\lambda}{\mu \cdot s \cdot (\mu - \lambda)} \cdot \left[\frac{1}{(1-p)} \right] [4].$$

Розглянемо приклад:

У великому супермаркеті є 4 каси, що працюють за системою М/М/с. Відомо, що середньоденна кількість покупців у магазині становить 480 клієнтів, а кожен касир обслуговує в середньому 20 клієнтів за годину. Каси працюють 8 годин на день. Оскільки наявна система часто призводить до довгих черг, керівництво хоче оптимізувати кількість кас, щоб зменшити час очікування клієнтів і покращити ефективність обслуговування.

Задача полягає в тому, щоб визначити оптимальну кількість кас за умов:

- Витрати на одного касира – 15 000 грн/місяць.
- Вартість втрат від очікування клієнтів – 100 грн/година.
- Бюджет на оптимізацію – 110 000 грн/місяць.

Розв'язання:

Потік клієнтів λ визначається як середня кількість клієнтів, які прибувають до системи за одиницю часу. Якщо в день магазин обслуговує 480 клієнтів за 8 годин, то потік клієнтів на годину буде:

$$\lambda = \frac{480}{8} = 60 \text{ клієнтів за годину.}$$

Кожен касир обслуговує 20 клієнтів на годину. Тому інтенсивність обслуговування на кожній касі:

$$\mu = 20 \text{ клієнтів на годину.}$$

Якщо для кожного варіанта ми обираємо кількість кас s , то загальні витрати на зарплату касирів за місяць будуть:

$$Z_{\text{зарплата}} = s \cdot 15\,000 \text{ грн/місяць.}$$

Загальні витрати на втрати через очікування:

$$Z_{\text{втрати}} = L_q \cdot W_q \cdot 100 \text{ грн/год.}$$

Для кожного варіанта кількості кас ссс будемо обчислювати загальні витрати Z_{total} , що включають витрати на зарплату касирів та витрати від втрат через очікування:

$$Z_{\text{total}} = Z_{\text{зарплата}} + Z_{\text{втрати}};$$

Нам потрібно вибрати таку кількість кас s , яка мінімізує Z_{total} , але водночас не перевищує бюджет у 110 000 грн/місяць.

Виконавши розрахунки, отримаємо таблицю з результатами.

Таблиця 1 – Результати обчислень

Кількість кас	Інтенсивність завантаження	Середній час очікування в черзі	Середня кількість клієнтів у черзі	Витрати на зарплату	Втрати через очікування	Загальні витрати
4	0,75	9	45	60 000	40 500	100 500
5	0,6	6	36	75 000	21 600	96 600
6	0,5	4,5	24	90 000	10 800	100 800
7	0,43	3,2	22,4	105 000	5 600	110 600

В результаті, для бюджету 110 000 грн/місяць оптимальним вибором буде 6 кас, оскільки загальні витрати становлять 100,800 грн/місяць, що є менше за бюджет.

Висновки. Отже, можна моделювати реальні процеси, аналізувати їх особливості та оптимізувати роботу систем за допомогою методів теорії черг. Параметри потоку клієнтів, рівень завантаженості серверів і час обслуговування можна точно оцінити за допомогою моделей М/М/1 і М/М/с. Використання теорії черг є важливою частиною процесу вдосконалення систем обслуговування, що сприяє розвитку сфери послуг і гарантує високий рівень задоволеності клієнтів.

Список використаних джерел

1. KA Group. URL: <https://kagroup.ua/tpost/iodsc6lfj1-teorya-cherh> (дата звернення: 25.11.2024).
2. Wikipedia. URL: https://en.wikipedia.org/wiki/Queueing_theory (дата звернення: 25.11.2024).
3. Science Direct. URL: <https://www.sciencedirect.com/topics/computer-science/queueing-theory> (дата звернення: 25.11.2024).
4. Real Statistics Using Excel. URL: <https://real-statistics.com/probability-functions/queueing-theory/m-m-s-queueing-model/> (дата звернення: 25.11.2024).