МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

ИССЛЕДОВАНИЕ ПРИМЕНИМОСТИ МЕТОДОВ НЕЧЕТКОГО ПОИСКА ДЛЯ ВЫЯВЛЕНИЯ СПЕЦИФИЧНОГО КОНТЕНТА НА ВЕБ-СТРАНИЦАХ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Киреевой Маргариты Николаевны

Научный руководитель:
доцент _____ Е.В. Кудрина

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Саратов 2023

ВВЕДЕНИЕ

Актуальность темы. С развитием интернета в нем появляется все больше контента, причем не всегда безопасного. Специфичный контент может быть просто контентом на определенную тематику, а может быть нежелательным или даже запрещенным для распространения. Так, веб-страницы могут содержать информацию о способах изготовления и использования наркотических средств, призывы к самоубийству, информацию о розничной дистанционной продаже алкоголя, информацию о способах самодельного изготовления различного оружия. То есть, контент, который содержит информацию, способную нанести вред человеку, считается нежелательным. К такому контенту относятся материалы, оказывающие негативное психологическое воздействие, и призывы к действиям или поведению, которые наносят ущерб человеку, группе лиц или самому себе. Часто нежелательным является контент, содержащий оскорбления, ругательства и нецензурные слова, элементы жестокости. Таким образом, в связи с тем, что подобного контента в интернете огромное количество, а опубликовать его довольно просто, проблема выявления специфичного контента очень актуальна.

Контентная фильтрация широко используется в образовании, в промышленности, госсекторе, правоохранительных органах, а также в органах госбезопасности. Поиск нежелательного и запрещенного контента — одна из составных частей фильтрации контента.

В целом, принцип работы любого контент-фильтра прост. Это программа, которая пропускает текст через себя и сравнивает различные последовательности символов с образцами, хранящимися в базе данных программы. Многие программы, реализующие поиск нежелательного контента, позволяют выполнять и другие действия (блокировки картинок, файлов, рекламы — DansGuardian, безопасный поиск — Entensys UserGate Web Filter, поиск утечек данных — Falcongaze SecureTower и т.д.).

Главной проблемой поиска специфичного контента является его скорость, так как необходимо обрабатывать большие объемы данных. Также проблематично то, что, как контент воспринимается, часто зависит от контекста. Специфичный контент можно искать по ключевым словам точно, однако такой поиск не учитывает опечатки в словах и их иные формы (множественное число, падеж). Последние проблемы могут решить методы нечеткого поиска, а наиболее хорошую скорость поиска можно выяснить сравнением различных методов. Поэтому в данной работе будет произведено исследование и сравнение методов нечеткого поиска специфичного контента.

Подобные исследования производились в работе [2], но в ней решалась задача классификации нежелательного контента. Также в работе [1] и [8] были рассмотрены некоторые методы выявления нежелательного контента, но без упоминания применения методов нечеткого поиска в данной задаче.

Цель бакалаврской работы – исследовать применимость методов нечеткого поиска для выявления специфичного контента на веб-страницах.

Поставленная цель определила **следующие задачи**:

1. Изучить понятие нечеткого поиска.
2. Сделать обзор подходов к нечеткому поиску специфического контента на веб-страницах.
3. Выбрать методы нечеткого поиска для более подробного изучения.
4. Сделать обзор инструментальных средств и технологий, используемых для реализации методов нечеткого поиска специфического контента в интернете.
5. Реализовать выбранные методы нечеткого поиска.
6. Выполнить сравнительный анализ эффективности поиска специфического контента реализованными методами.

Методологические основы исследования применимости методов нечеткого поиска для выявления специфичного контента на веб-страницах представлены в работах Галимовой А.Т. и Симонян А. Г [1], Чиркина Е.С. и Лопатина Д.В. [2] , Хуан Да [3], Сунь Ву и Уди Манбера [4].

Практическая значимость бакалаврской работы заключается в исследовании применимости методов нечеткого поиска, таких как алгоритм Bitap, метод N-грамм и алгоритм Вагнера-Фишера вычисления расстояния Левенштейна, для выявления специфичного контента на веб-страницах, их реализации, а также проведении сравнительного анализа реализованных методов.

Структура и объём работы. Бакалаврская работа состоит из введения, 2 разделов, заключения, списка использованных источников и 7 приложений. Общий объем работы – 60 страниц, из них 46 страниц – основное содержание, включая 13 рисунков и 17 таблиц, список использованных источников информации – 27 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретическая часть» посвящен постановке задачи нечеткого поиска, изучению специфики нечеткого поиска в интернете, обзору подходов к нечеткому поиску специфичного контента на веб-страницах и подробному изучению алгоритма Bitap, метода N-грамм и алгоритма Вагнера-Фишера вычисления расстояния Левенштейна.

В **подразделе 1.1 «Обзор подходов к нечеткому поиску специфического контента на веб-страницах»** рассматривается понятие нечеткого поиска, его специфика в интернете и различные методы нечеткого поиска.

Задача нечеткого поиска может быть поставлена следующим образом:

По заданному слову или фрагменту строки найти в тексте или словаре размера n все фрагменты строки, совпадающие с заданным с учетом k возможных различий.

Нечеткий поиск в интернете имеет свою специфику. Чтобы найти специфичный контент на исследуемых веб-страницах необходимо сначала собрать весь текст с них. Это можно сделать с помощью парсинга. Он

предназначен для сбора информации из Интернета и подготовки ее к автоматизированной обработке.

Далее текст необходимо привести к форме, приемлемой для сравнения. Это называется канонизацией или предобработкой текста.

Существуют разные алгоритмы нечеткого поиска специфичного контента на веб-странице. Кратко были рассмотрены следующие методы:

- анализ «множества слов» («bag of words», «мешок слов»);
- trie или префиксное дерево;
- фонетический поиск;
- хеширование текста и метод шинглов.

В **подразделах 1.2, 1.3, 1.4** рассматривались методы нечеткого поиска, выбранные для более подробного изучения и реализации. Это:

- алгоритм Bitap;
- метод N-грамм;
- алгоритм Вагнера-Фишера вычисления расстояния Левенштейна.

Алгоритм Bitap (также известный как shift-or, shift-and или алгоритм Баеза-Йейтса-Гоннета) начинается с предварительного вычисления набора битовых масок, содержащих по одному биту для каждого элемента шаблона. Затем он может выполнять большую часть работы с помощью побитовых операций, которые выполняются очень быстро.

Метод N-грамм — алгоритм, в котором сравнение слов осуществляется N-граммами — подстроками длины N. То есть, например, при поиске какого-либо слова оно разбивается на N-граммы и для каждой из них перебирается список слов, содержащих такую подстроку.

Расстояние Левенштейна — минимальное количество операций вставки, удаления или замены символа, с помощью которых можно превратить одну строку в другую. Расстояние Левенштейна вычисляет алгоритм Вагнера-Фишера.

Эти методы были выбраны, так как исследовать необходимо сильно отличающиеся друг от друга методы, например, метод N-грамм и Bitap используют совершенно разные средства для нечеткого поиска слова. Также вычисление расстояния Левенштейна является очень популярным методом, как и ранее упомянутые методы N-грамм и Bitap, что повлияло на выбор его в качестве исследуемого.

В **подразделе 1.5** производится обзор инструментальных средств и технологий, используемых для реализации методов нечеткого поиска специфического контента в интернете. Были рассмотрены следующие технологии:

- библиотеки Python для парсинга сайтов — Requests и BeautifulSoup;
- модули Python для вычисления времени выполнения определенного фрагмента программы — time, timeit и datetime module;
- библиотека Pandas на Python, которая используется для анализа данных и манипулирования данными;
- библиотеки визуализации на Python — matplotlib и Seaborn;
- метрики для оценки качества классификации и их функции из библиотеки scikit-learn, предназначенной для машинного обучения на языке Python, — accuracy, precision, recall и f-мера.

Были изучены сведения о Google Colab — среде, где производились все вычисления.

Итоги.

Были рассмотрены задача нечеткого поиска, ее специфика в интернете, различные подходы к нечеткому поиску специфичного контента на веб-страницах и выбраны для реализации и исследования алгоритм Bitap, метод N-грамм и алгоритм Вагнера-Фишера вычисления расстояния Левенштейна. Был сделан обзор необходимых инструментальных средств и технологий.

Второй раздел «Практическая часть» посвящен поиску необходимых данных, реализации методов нечеткого поиска: алгоритма Bitap, метода N-грамм и алгоритма Вагнера-Фишера вычисления расстояния Левенштейна, исследованию их применимости для выявления специфичного контента на веб-страницах и проведению сравнительного анализа реализованных методов.

В **подразделе 2.1** описывается найденный датасет со словами, разделенными по темам, его преобразование в словарь, посвященный теме растений, и составление списка веб-страниц, подходящих и не подходящих теме.

Так как нежелательный контент зачастую содержит нецензурные или другие оскорбляющие слова, по этическим соображениям было решено искать слова, относящиеся к определенной теме, которые, по сути, также являются специфическими. Был найден датасет, где каждому слову сопоставлен ровно один крупный тег (человек, предмет, действие). По этическим соображениям были взяты слова тега PLANT, то есть полностью относящиеся к теме растений. Размер полученного словаря — 251 слово.

Также для исследования необходимы сами веб-страницы, на которых будет производиться поиск специфических слов. Всего было самостоятельно собрано 100 веб-страниц, где 62 веб-страницы в датасете подходят теме растений, а 38 — нет.

Подраздел 2.2 посвящен подробному описанию реализации алгоритма Bitap и исследованию его применимости для выявления специфичного контента на веб-страницах.

Поиск специфических слов на веб-страницах с помощью алгоритма Bitap осуществляется следующим образом.

В цикле for с помощью парсинга собирается весь текст с каждой веб-страницы, для него производится предобработка, он разбивается на слова и каждое слово из словаря сравнивается со словом из текста алгоритмом Bitap с указанием максимально возможного числа ошибок. Если количество

ошибок, возвращенное методом, меньше максимально возможного числа ошибок, то число слов, соответствующих теме, увеличивается на единицу.

После подсчета количества слов, оно сравнивается с длиной текста на веб-странице.

Сравнение с полной длиной текста приводит к плохим результатам. Поэтому поделим длину текста на некоторое значение. Методом подбора было выбрано значение: длина текста, деленная на 250, так как при нем достигается максимальная точность метода. В результате сравнения переменной match, обозначающей соответствие теме, присваивается значение 0 или 1.

Был произведен перебор максимально возможного числа ошибок. При его нулевом значении, то есть при точном поиске, метод показал наиболее хорошие результаты — 93% точности. При значении этого числа 1 и 2 метод показал 68% точности.

Таким образом, алгоритм Bitap хорошо применим к поиску специфического контента на веб-странице только при его сведении к точному поиску.

Подраздел 2.3 посвящен подробному описанию реализации метода N-грамм и исследованию его применимости для выявления специфичного контента на веб-страницах.

Поиск специфических слов на веб-страницах осуществляется похоже на поиск с помощью алгоритма Bitap, но есть значительные отличия. Сначала вызывается функция создания словаря. Также в цикле for с помощью парсинга производятся аналогичные действия, но теперь каждое слово из текста ищется в составленном словаре N-грамм.

Также перебиралось значение для сравнения количества слов, подходящих теме, с длиной текста. Время выполнения метода составляет 133 секунды.

Несмотря на то, что наиболее часто используют триграммы, при длине N-грамм, равной 3, точность составила только 0.68, а при длине, равной 4,

уже 0.98. При длине, равной 3, метод находил много лишних слов, неподходящих теме. Так, зачастую проблема в том, что метод в длинных словах из текста находит короткие из словаря. Если увеличить длину N-граммы, то слова из трех букв, которых достаточно много в словаре, метод в длинных словах из текста не будет находить. При длине N-грамм, равной 4, таких слов значительно меньше, поэтому эта длина является оптимальной для нашей задачи.

Таким образом, можно сказать, что метод N-грамм хорошо применим для выявления специфического контента.

Подраздел 2.4 посвящен описанию реализации алгоритма Вагнера-Фишера вычисления расстояния Левенштейна и исследованию его применимости для выявления специфичного контента на веб-страницах.

Поиск специфических слов на веб-страницах осуществляется похоже на поиск с помощью алгоритма Bitap. Изменено только условие увеличения количества слов, соответствующих теме: если вычисленное расстояние Левенштейна меньше длины слова из текста, деленной на 3 (это значение было подобрано), то число слов, соответствующих теме, увеличивается на единицу.

Время выполнения данного метода составило 1177 секунд, а его точность равна 1, поэтому метод хорошо применим для выявления специфического контента.

Подраздел 2.5 посвящен сравнительному анализу реализованных ранее методов. В таблице 1 представлены точность и время выполнения методов при переборе 100 веб-страниц из собранного ранее датасета.

Таблица 1 – Время выполнения и точность всех методов с указанием их параметров

Метод	Параметры метода	Время выполнения	Точность (от 0 до 1)
Bitap	максимальное количество ошибок = 0	603 с	0.93
Bitap	максимальное количество ошибок = 1	803 с	0.68
Bitap	максимальное количество ошибок = 2	1003 с	0.68
Bitap	максимальное количество ошибок = 3	1174 с	0.62
N-грамм	длина = 3	133 с	0.68

N-грамм	длина = 4	133 с	0.98
Алгоритм Вагнера-Фишера (расстояние Левенштейна)	нет особых параметров	1177 с	1.00

Проанализировав данные, представленные в таблице 1, можно сделать следующие выводы.

Так как перебор веб-страниц с использованием метода N-грамм содержит всего один цикл, в отличие от перебора с использованием других алгоритмов, можно предположить, что он будет значительно быстрее. Так и есть, его время выполнения всего 133 секунды. Также здесь важную роль играет небольшой размер словарей, используемых в методе.

Алгоритм Bitap не вычисляет расстояние Левенштейна полностью, поэтому он быстрее алгоритма Вагнера-Фишера, который это делает. Также алгоритм Bitap использует побитовые операции, которые должны работать быстрее арифметических.

Максимально точным оказался алгоритм Вагнера-Фишера вычисления расстояния Левенштейна, скорее всего, из-за того, что он как раз сравнивает все слова полностью между друг другом.

К алгоритму Вагнера-Фишера близок по точности метод N-грамм, но при длине 3 он также может найти достаточно много слов, не подходящих теме. Проблема в том, что метод в длинных словах из текста находит много коротких из словаря. При длине 4 таких слов мало, поэтому точность получается хорошей.

Алгоритм Bitap менее точен по сравнению с предыдущими методами. Хорошие результаты он показывает только при нулевом количестве ошибок, допустимых при поиске, то есть при сведении поиска к точному. Проблема здесь получается похожа на проблему из метода N-грамм.

Также были вычислены другие оценки качества — f-мера и ее составные: precision и recall. Результаты с помощью f-меры отличаются лишь значением, порядок отсортированных по точности методов сохранился.

Итоги.

Алгоритм Bitap хуже всего из рассмотренных методов нечеткого поиска применим к задаче отнесения контента на веб-страницы к определенной теме.

Метод N-грамм работает и быстро, и почти идеально точно при длине N-грамм, равной 4. Алгоритм Вагнера-Фишера вычисления расстояния Левенштейна работает медленно, но зато может выдать максимальную точность. Соответственно, при выборе метода для выявления специфического контента на веб-странице нужно ориентироваться, важна ли скорость метода и насколько большой объем данных нужно обработать.

ЗАКЛЮЧЕНИЕ

В ходе выполнения бакалаврской работы были решены все поставленные задачи, а именно изучено понятие нечеткого поиска и сделан обзор подходов к нечеткому поиску специфического контента на веб-страницах. Для более подробного изучения и реализации были выбраны алгоритм Bitap, метод N-грамм и алгоритм Вагнера-Фишера вычисления расстояния Левенштейна. Был сделан обзор инструментальных средств и технологий, используемых для реализации методов нечеткого поиска специфического контента в интернете. Далее методы были реализованы и между ними был проведен сравнительный анализ по их времени выполнения и оценке качества. Сравнительный анализ показал, что наилучший результат для поиска специфичного контента на веб-сайтах за приемлемое время показывает метод N-грамм, однако если время не критично, то лучше использовать метод Вагнера-Фишера нахождения расстояния Левенштейна. Таким образом, цель бакалаврской работы достигнута.

Следует отметить, что исследование методов нечеткого поиска проводилось на основе одной специфической темы «растения» датасета, содержащего результаты разметки слов и выражений русского языка по семантическим срезам. Выбор темы был обусловлен этическими

соображениями. Поменяв специфику датасета, можно применить рассмотренные методы нечеткого поиска для выявления нежелательного и/или запрещенного контента на веб-сайтах с целью дальнейшей контент-фильтрации.

В дальнейшем, тему бакалаврской работы можно развить до полной классификации веб-страниц по темам.

Отдельные части бакалаврской работы были опубликованы:

1. Киреева М.Н. Сравнительный анализ методов вычисления редакционного расстояния: программная реализация и их оптимизация // Фундаментальные и прикладные аспекты развития современной науки / Сборник научных статей по материалам X Международной научно-практической конференции, с.330-336 г. Уфа, 17 января 2023г. URL: https://perviy-vestnik.ru/archive-konferentsiya-2023/ (дата обращения: 05.05.2023).
2. Киреева М.Н. Программная оптимизация алгоритма Вагнера-Фишера для вычисления редакционного расстояния//Студенческий вестник: электрон. научн. журн. 2022. №20(212) URL: https://studvestnik.ru/journal/stud/herald/212 (дата обращения: 05.05.2023).

Основные источники информации:

1. Галимова, А. Т. Методы выявления нежелательного контента в тексте и изображениях / А. Т. Галимова, А. Г. Симонян // Технологии информационного общества: Сборник трудов XV Международной отраслевой научно-технической конференции «Технологии информационного общества», Москва, 03–04 марта 2021 года. – Москва: ООО "Издательский дом Медиа паблишер", 2021. – С. 151-152. URL: https://www.elibrary.ru/item.asp?id=45671294 (дата обращения 11.05.2023)
2. Чиркин Е.С., Лопатин Д.В. Подходы к нечеткому поиску нежелательного контента на веб-странице // Вестник Тамбовского университета. Серия Естественные и технические науки. Тамбов, 2016. Т. 21. № 6. С. 2358-

2365. DOI: 10.20310/1810-0198-2016-21- 6-2358-2365. URL: http://journals.tsutmb.ru/a8/upload/2017-end/2358-2365_—ËaЄË%C2%AD__‹®Ї%C2%A0вË%C2%AD.pdf (дата обращения 08.11.2022)

3. Хуан Д. Алгоритмы извлечения информации из текстов, парсинг веб-страниц с использованием языка программирования Python // Актуальные исследования. 2022. №30 (109). С. 21-24. URL: https://apni.ru/article/4419-algoritmi-izvlecheniya-informatsii-iz-tekstov (дата обращения: 11.05.2023)
4. Sun Wu and Udi Manber. 1992. Fast text searching: allowing errors. Commun. ACM 35, 10 (Oct. 1992), 83–91. URL: https://doi.org/10.1145/135239.135244 (дата обращения 12.11.2022)
5. Нечёткий поиск в тексте и словаре [Электронный ресурс] URL: https://habr.com/ru/post/114997/ (дата обращения 27.10.2022)
6. Задача о редакционном расстоянии, алгоритм Вагнера-Фишера — Викиконспекты кафедры компьютерных технологий Университета ИТМО [Электронный ресурс] URL: https://neerc.ifmo.ru/wiki/index.php?title=Задача_о_редакционном_расстоянии,_алгоритм_Вагнера-Фишера (дата обращения 12.11.2022)
7. Алгоритм Shift-And — Викиконспекты кафедры компьютерных технологий Университета ИТМО [Электронный ресурс] URL: https://neerc.ifmo.ru/wiki/index.php?title=Алгоритм_Shift-And (дата обращения 10.11.2022)
8. Миюзов, Р. Е. Применение технологии искусственного интеллекта при модерации контента в сети Интернет / Р. Е. Миюзов // СТУДЕНЧЕСКАЯ НАУКА: АКТУАЛЬНЫЕ ВОПРОСЫ, ДОСТИЖЕНИЯ и ИННОВАЦИИ: сборник статей Международной научно-практической конференции, Пенза, 17 апреля 2021 года. – Пенза: Наука и Просвещение, 2021. – С. 41-45. URL: https://www.elibrary.ru/item.asp?id=45702774 (дата обращения 11.05.2023)