

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ
ФЕДЕРАЦИИ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

Национальный исследовательский ядерный университет
“МИФИ”

Факультет Кибернетики и информационной безопасности
Кафедра Информационные технологии в социальных системах

К защите допущен
Зам. заведующего кафедрой
_____ /Сергиевский М. В./
“_____” _____ 201__ г.

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
к учебно-исследовательской работе
и курсовому проекту
на тему:

**РАЗРАБОТКА АЛГОРИТМА ДЛЯ ВЫДЕЛЕНИЯ
ЧАСТОВСТРЕЧАЮЩИХСЯ ШАБЛОНОВ ОШИБОК ИЗ
ФАЙЛОВ ЖУРНАЛОВ ПРИЛОЖЕНИЙ**

Студент:

Тропин А. Г.

Руководитель

Сергиевский М. В.

Оценка

Подпись членов комиссии

Москва 2015

РЕФЕРАТ

Пояснительная записка к учебно-исследовательской работе содержит 21 страниц.

Ключевые слова: `mapreduce`, `logs`, распределенные вычисления, `python`, надежность, отказоустойчивость, функциональное программирование, регулярные выражения.

Цель работы — разработка алгоритма для выделения частовстречающихся шаблонов ошибок из файлов журналов приложений.

В процессе работы осуществлялась разработка алгоритма, реализация алгоритма на языке `python`. И внедрение в производство с использованием MapReduce-системы Yandex Tables.

В результате работы был разработан алгоритм, реализован на языке `python` и написаны две вспомогательные утилиты на языке `bash` и две утилиты на языке `python`.

СОДЕРЖАНИЕ

Реферат	1
Нормативные ссылки	3
Определения, обозначения и сокращения	4
Введение	5
1 Современное состояние проблемы анализа файлов журналов	6
1.1 Постановка задачи	6
1.2 Актуальность задачи	6
1.3 Сравнение с существующими аналогами	7
2 Проектирование алгоритма	8
2.1 Структура	8
2.2 Архитектура	10
2.3 Планируемая выгода	10
3 Реализация	11
3.1 Сравнение возможных решений	11
3.2 Особенности реализации, проблемы и способы их решения .	11
3.3 Алгоритм работы программы	11
4 Тестирование и анализ результатов	12
4.1 Методы тестирования	12
4.2 Анализ результатов тестирования	12
4.3 Дальнейшее развитие	12
Заключение	13
Список использованных источников	14
Приложения	15

НОРМАТИВНЫЕ ССЫЛКИ

В настоящей пояснительной записке к учебно-исследовательской работе использованы ссылки на следующие стандарты.

ГОСТ 7.32-2001 Система стандартов по информации, библиотечному и издательскому делу. Отчет о научно-исследовательской работе. Структура и правила оформления.

ГОСТ 7.9-95 Система стандартов по информации, библиотечному издательскому делу. Реферат и аннотация. Общие требования.

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

Определение 1 (MapReduce). Модель распределённых вычислений, представленная компанией Google, используемая для параллельных вычислений над очень большими, несколько петабайт, наборами данных в компьютерных кластерах.

Определение 2 (Map). $\text{map}(f, \text{list})$ — функция высшего порядка двух аргументов, применяет к каждому элементу списка list , функцию $f(x)$, в качестве результата возвращает список полученных значений.

Пример. $\text{map}(\text{lambda } x: x**3, [1, 2, 3])$ вернёт список $[1, 8, 27]$.

Определение 3 (Reduce). $\text{reduce}(f, \text{list}, \text{init}=\text{None})$ — функция высшего порядка, последовательно применяет $f(x, y)$ к элементу списка и значению от предыдущего выполнения функции.

Пример. $\text{reduce}(\text{lambda } x, y: x + y, [1, 2, 3])$ вернёт 6 (сумма всех элементов).
 $6 = \text{sum}(\text{sum}(1, 2), 3)$.

Определение 4 (Экземпляр). Приложение, запущенное в контейнере и описываемое парой `host:port`.

Определение 5 (Шаблон). Специально подготовленное регулярное выражение с экранированными спецсимволами.

Определение 6 (Эвристический алгоритм (эвристика)). Алгоритм решения задачи, не имеющий строгого обоснования, но, тем не менее, дающий приемлемое решение задачи в большинстве практически значимых случаев.

ВВЕДЕНИЕ

Большинство программных систем, имеющих сложную структуру и состоящих из нескольких сотен различных компонент, обладают рядом схожих проблем. Например веб-поиск содержит следующие компоненты: балансеры, верхние, средние метапоиски, промежуточные и базовые поиски, колдунщики, антироботы, свежесть, региональные поиски, несколько десятков параллельных поисков.

Всего одновременно запущено несколько ***** тысяч экземпляров приложений. Каждый экземпляр генерирует множество сообщений об ошибках и записывает каждое из них в файл журнала ошибок.

Некоторые приложения, близкие по функционалу, пишут в один и тот же файл. Файлы журналов ротируются согласно определённому алгоритму. Тем не менее объем файла журнала для одного экземпляра может достигать нескольких сотен мегабайт, что препятствует быстрому ручному анализу в случае инцидента и инженеры вынуждены тратить ценные секунды на просмотр сотен тысяч строк файла в поисках сообщения с описанием элемента, вызвавшего сбой работы системы.

Начальным требованием к системе для эффективного использования алгоритма является наличие сопоставимого с количеством поисковых экземпляров приложений количества узлов на которых может быть запущена программа, реализующая алгоритм.

В организации, в которой выполнялась учебно-исследовательская работа, развёрнута большая поисковая инфраструктура, которая не лишена недостатков и существует вероятность поломки некоторой её части. Существует множество средств мониторинга состояния веб-поиска и противодействия инцидентам, но в некоторых случаях инженерам их недостаточно и приходится вручную анализировать файлы журналов отдельных экземпляров приложений на отдельных серверах, что, в свою очередь, замедляет скорость реакции на непредвиденную ситуацию. Но даже автоматизация процесса анализа файла журнала одного экземпляра не решает проблему полностью, поэтому необходима возможность быстрого анализа файлов журналов сразу множества экземпляров.

Таким образом, целью этой учебно-исследовательской работы является разработка алгоритма, позволяющего собирать статистику по ошибкам, встречающимся в файлах журналов экземпляров поисковых приложений, на основе существующих шаблонов и выделять новые шаблоны.

1 СОВРЕМЕННОЕ СОСТОЯНИЕ ПРОБЛЕМЫ АНАЛИЗА ФАЙЛОВ ЖУРНАЛОВ

1.1 Постановка задачи

Необходимо разработать алгоритм, позволяющий:

- Собирать статистику с использованием известных шаблонов.
- Выделять новые шаблоны с помощью эвристических методов.
- Производить лёгкое масштабирование программы, реализующей алгоритм с использованием существующих технологий.

Требование к программе, реализующей алгоритм:

- Высокая производительность(файл, размеров в 100МБ должен обрабатываться не более 15 секунд).
- Возможность запуска под операционной системой семейства GNU/Linux.
- Горизонтальная масштабируемость.

1.2 Актуальность задачи

Достаточно крупных систем, для которых поставленная задача могла бы иметь актуальность, не так много и в основном это коммерческие проекты таких компаний как: google, yahoo, yandex, amazon, facebook, twitter. Откуда вытекает два обстоятельства:

- а) Продукты разрабатываемые компаниями не выкладываются в свободный доступ.
- б) Продукты разрабатываемые сторонними компаниями не учитывают специфику крупных распределённых систем.

Ещё одним важным аспектом является тот факт, что существующие приложения предполагают, что формат файла журнала заранее известен, но это бывает не всегда так. Например, если продукт разрабатывается в течении 20 лет исправление ошибки, заложенной в начале разработки, обходится очень дорого. Поэтому на порядок дешевле написать алгоритм, который позволяет анализировать неструктурированные файлы журналов, чем исправлять архитектурные ошибки.

1.3 Сравнение с существующими аналогами

Ниже представлен список существующих проприетарных и свободных решений, выложенных в открытом доступе, с описанием их недостатков:

- ascolog. <https://www.ascolog.com/content/about-ascolog>
 - Проприетарный. Отсутствие исходного кода создаёт множество проблем, таких как: отсутствие возможности расширения функционала, возможное наличие вредоносных компонентов.
 - Необходимая операционная система для запуска: Windows. К сожалению это достаточно редкое решение для крупных распределённых систем. Критичный пункт.
 - Нет возможности для горизонтального масштабирования. Вертикальное масштабирование сильно ограничено по своим возможностям. Ещё один критичный пункт.
 - Нет возможности для расширения функционала, какими-либо встроенными средствами.
 - Сайт производителя не обновлялся больше 6 месяцев и, по всей видимости, продукт больше не развивается.
- sawmill. <https://www.sawmill.net/index.html>
 - Проприетарный. Отсутствие исходного кода создаёт множество проблем, таких как: отсутствие возможности расширения функционала, возможное наличие вредоносных компонентов.
 - Нет возможности для расширения функционала, какими-либо встроенными средствами.
 - Предполагается заранее известная структура логов. Критичный пункт.

Из-за наличия критичных проблем каждого из вышеперечисленных решений, они не подходят для решения поставленной задачи.

2 ПРОЕКТИРОВАНИЕ АЛГОРИТМА

2.1 Структура

Структура входных данных(файла журнала) разнородна. Это может быть как стек вызовов функции языка javascript(несколько строк). Так и ошибки, вызова функций языка perl. Так и отладочные сообщения о входящем запросе неправильного формата. Все эти сообщения имеют принципиально разный формат, к тому же некоторые имеют мета-информацию(временная отметка, важность и т.д.), другие — нет.

За счёт этого, даже сообщения об одной и той же ошибке могут совпадать неполностью(различные временные отметки) или практически полностью несовпадать — информация об ошибке и длинный путь до файла.

Поэтому на первом этапе необходимо унифицировать сообщения об ошибках за счёт замены мета-информации и дополнительных сведений на строковые константы:

Пример (Замена мета-информации и доп. сведений). UUID вида ACE088EB-ECA6-4348-905A-041EF10DBD53 можно заменить на строковую константу вида UUID_SUBSTITUTE.

Пример (Замена мета-информации и доп. сведений). ip-адрес вида 10.205.10.74 можно заменить на строковую константу вида IP_ADDRESS_SUBSTITUTE.

После этого данные станет проще анализировать с помощью алгоритмов, так как одинаковые ошибки будут совпадать, независимо от времени появления или идентификатора запроса.

Для поиска и замены удобно использовать регулярные выражения, например:

- '(\d{16}-[-\w]*\b)', 'REQUEST_ID_SUBSTITUTE'
- '([0-9A-F]){8}-[0-9A-F]{4}-[0-9A-F]{4}-[0-9A-F]{4}-[0-9A-F]{12}', 'UUID_SUBSTITUTE'
- '\b(\d{1,3}\.){3}\d{1,3}\b', 'IP_ADDRESS_SUBSTITUTE'
- '\w{3} \w{3} \d{1,2} \d{1,2}:\d{1,2}:\d{1,2} \d{4}', 'TIMESTAMP_SUBSTITUTE'
- 'line \d+', 'LINE_SUBSTITUTE'

Такого рода пары(регулярное выражение, строковая константа) удобно хранить и использовать. Регулярные выражения — гибкий инструмент, поддерживаемый практически всеми современными языками программирования.

Сами сообщения об ошибках можно представить в виде регулярных выражений. Но зачем тогда делать первый этап преобразования входных данных? Теоретически можно описать сообщение об ошибке регулярным выражением, но на практике это регулярное выражение будет достаточно сложным и нечитаемым, поэтому подготовка начальных данных сильно упрощает регулярные выражения, используемые для выделения ошибки. Например следующий многострочный шаблон было бы воспринимать человеку ещё сложнее, если бы, вместо строковых констант были бы регулярные выражения:

```

""JavaScript\\.\\ TypeError\\:\\ Object\\
\\#\\<t\\>\\ has\\ no\\ method\\ \\'Ukraine\\'
\\ \\ \\ \\ at\\ Object\\.blocks\\.b\\-counters\\
\\_\\_gemius\\ \\(web3\\_exp\\/pages\\-desktop\\/
search\\/\\_search\\.all\\.priv\\.js\\:POSITION\\_SUBSTITUTE\\)
\\ \\ \\ \\ at\\ Object\\.\\.\\[object\\ Function\\]\\
\\.Object\\.toString\\.call\\.e\\.\\.\\(anonymous\\ function\\)\\
\\[as\\ b\\-counters\\_\\_gemius\\]\\ \\(web3\\_exp\\/pages\\
-desktop\\/search\\/\\_search\\.all\\.priv\\.js\\:POSITION\\_SUBSTITUTE\\)
\\ \\ \\ \\ at\\ Object\\.blocks\\.b\\-counters\\.blocks\\.b\\-cntrs\\
\\(web3\\_exp\\/pages\\-desktop\\/search\\/\\_search\\.all\\
\\.priv\\.js\\:POSITION\\_SUBSTITUTE\\)
\\ \\ \\ \\ at\\ Object\\.\\.\\[object\\ Function\\]\\.Object\\
\\.toString\\.call\\.e\\.\\.\\(anonymous\\ function\\)\\
\\[as\\ b\\-counters\\]\\ \\(web3\\_exp\\/pages\\-desktop\\/
search\\/\\_search\\.all\\.priv\\.js\\:POSITION\\_SUBSTITUTE\\)
\\ \\ \\ \\ at\\ Object\\.\\.\\[object\\ Function\\]\\.Object\\.toString\\.
call\\.e\\.\\.\\(anonymous\\ function\\)\\ \\[as\\ b\\-page\\]\\
\\(web3\\_exp\\/pages\\-desktop\\/search\\/\\_search\\.all\\
\\.priv\\.js\\:POSITION\\_SUBSTITUTE\\)\\ at\\
\\/db\\/BASE\\/upper\\-SHARD\\_SUBSTITUTE\\/arkanavt\\/report\\
\\/lib\\/YxWeb\\/Util\\/Template\\/JS\\.pm\\ LINE\\_SUBSTITUTE\\.\\
\\(Object\\.blocks\\.b\\-page\\ \\(web3\\_exp\\/pages\\-desktop\\/search\\/
\\_search\\.all\\.priv\\.js\\:POSITION\\_SUBSTITUTE\\)\\)\\n\\Z""",

```

Также использование дополнительных регулярных выражений в шаблонах, приводит к падению производительности, увеличивает вероятность ошибки при подсчёте статистики и усложняет автоматическое выделение новых шаблонов.

Для подсчёта статистики используется достаточно простая структура — пара (шаблон, количество совпадений). Это один из ключевых моментов, позволяющих легко масштабировать приложение, реализующее алгоритм, с помощью модели MapReduce.

2.2 Архитектура

2.3 Планируемая выгода

3 РЕАЛИЗАЦИЯ

3.1 Сравнение возможных решений

3.2 Особенности реализации, проблемы и способы их решения

про `collet-logs.py` и `skynet`

3.3 Алгоритм работы программы

4 ТЕСТИРОВАНИЕ И АНАЛИЗ РЕЗУЛЬТАТОВ

4.1 Методы тестирования

4.2 Анализ результатов тестирования

4.3 Дальнейшее развитие

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

ПРИЛОЖЕНИЯ

```
#!/usr/bin/env python
# encoding: utf-8

import sys
import os
import re
import argparse
import fileinput
import copy
import itertools

from argparse import RawDescriptionHelpFormatter

from patterns import main_patterns

BUFFER_SIZE = 30

class Buffer(object):
    """Class for buffering input"""

    def __init__(self, output_filename=os.devnull):
        self.buf = [] * BUFFER_SIZE
        self.line_scrolled = 0
        self.matched_lines = []
        self.unmatched_stream = open(output_filename, 'w')
        self.updated = True
        self.last_text = ''

    def add(self, line):
        self.buf.append(line)
        if 0 not in self.matched_lines:
            self.unmatched_stream.write(line)

        self.buf = self.buf[1:BUFFER_SIZE+1]
        self.matched_lines = filter(lambda x: x >= 0,
                                     map(lambda x: x - 1, self.matched_lines))
        self.line_scrolled += 1
        self.updated = True

    def text(self):
        """lazy function, returning joined lines
        :returns: str()

        """
        if self.updated:
            self.last_text = ''.join(self.buf)
            self.updated = False

        return self.last_text
```



```

def try_to_match(self, pattern):
    matched = pattern.search(self.text())
    if matched:
        self.mark_matched(pattern)
    return matched

def mark_matched(self, pattern):
    lines_count = len(pattern.split('\n'))
    marked_lines = range(BUFFER_SIZE - lines_count, BUFFER_SIZE)
    self.matched_lines += marked_lines

def __enter__(self):
    return self

def __exit__(self, type, value, traceback):
    for _ in xrange(BUFFER_SIZE):
        self.add('')
    self.unmatched_stream.close()

class SnippetsQueue(object):
    def __init__(self):
        self.new_snippets = []
        self.ready_snippets = {}
        self.pattern_used = {}

    def push(self, snippet):
        if not self.pattern_used.get(pattern, False):
            self.new_snippets.append(snippet)
            self.pattern_used[pattern] = True

    def add(self, line):
        for snippet in self.new_snippets:
            if snippet.full():
                self.make_ready(snippet)
            else:
                snippet.add(line)

    def make_ready(self, snippet):
        self.new_snippets.remove(snippet)
        pattern = snippet.pattern
        try:
            self.ready_snippets[pattern].append(snippet)
        except KeyError:
            self.ready_snippets[pattern] = [snippet]
        self.pattern_used[pattern] = False
        if len(self.ready_snippets[pattern]) == SNIPPETS_TO_SHOW:
            self.pattern_used[pattern] = True

    def make_all_ready(self):
        for snippet in self.new_snippets:

```

```

        self.make_ready(snippet)

class StatCollector(object):

    """Class for collecting statistics"""

    def __init__(self):
        self.match_count = {}
        self.lines_count = 0
        self.number_of_matched_lines = 0
        pass

    def add(self, pattern):
        """Add pattern matching event"""
        if pattern not in self.match_count:
            self.match_count[pattern] = 0
        self.match_count[pattern] += 1
        self.number_of_matched_lines += len(pattern.split('\n'))

def print_stat(stat_collector, snippets_queue=None, **kwargs):

    """Print statistics and snippets"""

    for pattern, count in stat_collector.match_count.iteritems():
        print "\n\
*****\n\
pattern: {}"

        -----

        number of matches: {}
        *****\
        """.format(pattern, count)
        if snippets_queue:
            if pattern in snippets_queue.ready_snippets:
                for snippet in snippets_queue.ready_snippets[pattern]:
                    snippet.show()
                    output_stream.write('-' * 80 + '\n')
            else:
                output_stream.write('|No snippets found : (\n')
                output_stream.write('-' * 80 + '\n')

    print 'patterns used: {}/{}'.format(
        len(stat_collector.match_count),
        len(kwargs['patterns']))
    print 'number of lines matched: {}/{}'.format(
        stat_collector.number_of_matched_lines,
        stat_collector.lines_count)

def make_escaped(string):

```

```

"""Make escaped pattern from string

:string: string to be escaped
:returns: pattern

"""

return re.escape(string.replace('\n', 'NSLPH')).replace('NSLPH', '\n') + r'\n\Z'

def get_stat(input_files, snippets_count=0, context=3,
             patterns=main_patterns,
             output_stream=sys.stdout, snippets_file=None,
             unmatched_filename=os.devnull):
    """Show statistics for patterns

    :input_stream: input stream
    :snippets_count: maximum number of snippets to show for every pattern
    :context: number of lines in snippet around last line of pattern match
    :patterns: patterns that need to look
    :output_stream: stream for output
    :returns: None

    """

    input_stream = fileinput.input(input_files)

    LINES_ABOVE = context
    LINES_BELOW = context
    SNIPPETS_TO_SHOW = snippets_count
    SHOW_SNIPPETS = snippets_count > 0

    class Snippet(object):
        def __init__(self, lines_above, pattern, line_number):
            self.text = copy.copy(lines_above)
            self.pattern = pattern
            self.line_number = line_number

        def full(self):
            return len(self.text) >= LINES_ABOVE + LINES_BELOW + 1

        def add(self, line):
            self.text.append(line)

        def show(self):
            i = 0
            for line in self.text:
                if i == LINES_ABOVE:
                    output_stream.write('| {} ==>'.format(self.line_number))
                else:
                    if re.search(self.pattern, line):
                        output_stream.write('-->')

```

```

        else:
            output_stream.write('|>')
        output_stream.write(line)
        i += 1

stat_collector = StatCollector()
snippets_queue = SnippetsQueue()
line_number = 1

compiled_patterns = map(re.compile, patterns)

with Buffer(unmatched_filename) as input_buffer:
    if snippets_file:
        snippets_buffer = Buffer()
    else:
        snippets_buffer = input_buffer

    for line in input_stream:
        input_buffer.add(line)
        if snippets_file:
            line = snippets_file.readline()
            snippets_buffer.add(line)

        for pattern in compiled_patterns:
            if input_buffer.try_to_match(pattern):
                if SHOW_SNIPPETS:
                    if LINES_ABOVE > 0:
                        snippet_beginning = snippets_buffer.buf[-LINES_ABOVE:]
                    else:
                        snippet_beginning = []

                    snippet = Snippet(snippet_beginning, pattern.pattern,
                                      line_number)
                    snippets_queue.push(snippet)
                    stat_collector.add(pattern.pattern)

                if SHOW_SNIPPETS:
                    snippets_queue.add(line)
                    line_number += 1

stat_collector.lines_count = line_number - 1

if not SHOW_SNIPPETS:
    snippets_queue = None
else:
    snippets_queue.make_all_ready()

return (stat_collector, snippets_queue)

def show_patterns():

```

```

"""Show patterns from patterns.py"""

for pattern in main_patterns:
    print pattern
    print '-' * 80

def print_escaped(files):
    input_stream = fileinput.input(files)
    text = ''
    for line in input_stream:
        text += line
    escaped_text = make_escaped(text)
    print '\n'
    print escaped_text

def main():
    parser = argparse.ArgumentParser(description=\
    """
        Parse file[s]\n\n
        examples:
        ./%(prog)s file[s] -s 2 -C 3
        cat error_log | tail -n 1000 | ./prepare.py | ./%(prog)s
        ./%(prog)s -p
        echo "some \\ntext, that wants to be pattern" | ./%(prog)s -e
    """, formatter_class=RawDescriptionHelpFormatter)
    parser.add_argument('file', nargs='*', default=[],
                        help='file[s] to be parsed')
    parser.add_argument('-s', '--snippets', nargs='?', type=int, const=5,
                        help='show maximum SNIPPETS snippets (5 default)')
    parser.add_argument('-C', '--context', nargs='?', type=int,
                        help='show CONTEXT lines around pattern last line')
    parser.add_argument('-p', '--patterns', action='store_const', const=True,
                        help='show patterns')
    parser.add_argument('-e', '--escape', action='store_const', const=True,
                        help='escape given string')
    parser.add_argument('-o', '--original', nargs=1,
                        help='provide original file for better snippets')
    parser.add_argument('-u', '--unmatched', nargs=1,
                        help='output unmatched text to file')

    args = parser.parse_args()

    if args.patterns:
        show_patterns()
        return

    if args.escape:
        print_escaped(args.file)
        return

```

```

kwargs = {}

kwargs['input_files'] = args.file

if args.snippets:
    kwargs['snippets_count'] = args.snippets

if args.context is not None:
    kwargs['context'] = args.context

if args.original:
    kwargs['snippets_file'] = fileinput.input(args.original)

if args.unmatched is not None:
    kwargs['unmatched_filename'] = args.unmatched[0]

kwargs['patterns'] = main_patterns

result = get_stat(**kwargs)
print_stat(*result, **kwargs)

if __name__ == '__main__':
    main()

```