



**Politechnika
Śląska**

PRACA MAGISTERSKA

Analiza narzędzi wyszukiujących zawartość tekstową w systemie Linux

Kacper NITKIEWICZ

Nr albumu: 290409

Kierunek: **⟨wpisać właściwy⟩**

Specjalność: **⟨wpisać właściwą⟩**

PROWADZĄCY PRACĘ

Dr inż. Adrian Smagór

KATEDRA **⟨wpisać właściwą⟩**

Wydział Automatyki, Elektroniki i Informatyki

OPIEKUN, PROMOTOR POMOCNICZY

⟨stopień naukowy imię i nazwisko⟩

Gliwice 2024

Tytuł pracy

Analiza narzędzi wyszukiwujących zawartość tekstową w systemie Linux

Streszczenie

(Streszczenie pracy – odpowiednie pole w systemie APD powinno zawierać kopię tego streszczenia.)

Słowa kluczowe

(2-5 słów (fraz) kluczowych, oddzielonych przecinkami)

Thesis title

Thesis title in English

Abstract

(Thesis abstract – to be copied into an appropriate field during an electronic submission – in English.)

Key words

(2-5 keywords, separated by commas)

Spis treści

1	Wstęp	1
2	Analiza tematu wyszukiwania tekstu	3
2.1	Sformułowanie problemu	3
2.2	State of art	4
2.3	Opis poznanych rozwiązań	4
2.3.1	Algorytm brute force	4
3	[Tytuł rozdziału]	9
3.1	[Tytuł podrozdziału]	9
3.2	[Tytuł podrozdziału]	9
4	Badania	11
5	Podsumowanie	13
	Bibliografia	15
	Dokumentacja techniczna	19
	Spis skrótów i symboli	21
	Lista dodatkowych plików, uzupełniających tekst pracy (jeżeli dotyczy)	23
	Spis rysunków	25
	Spis tabel	27

Rozdział 1

Wstęp

Rozdział 2

Analiza tematu wyszukiwania tekstu

$$y = \frac{\partial x}{\partial t} \tag{2.1}$$

- analiza tematu
- wprowadzenie do dziedziny (*state of the art*) – sformułowanie problemu,
- poszerzone studia literaturowe, przegląd literatury tematu (należy wskazać źródła wszystkich informacji zawartych w pracy)
- opis znanych rozwiązań, algorytmów, osadzenie pracy w kontekście
- Tytuł rozdziału jest często zbliżony do tematu pracy.
- Rozdział jest wysycony cytowaniami do literatury [3, 4, 2]. Cytowanie książki [4], artykułu w czasopiśmie [3], artykułu konferencyjnego [2] lub strony internetowej [1].

2.1 Sformułowanie problemu

Wyszukiwanie tekstu w systemach towarzyszy ludziom od początków istnienia maszyn, choć pierwsze komputery nie posiadały ogromnych ilości pamięci co nie powodowało potrzeby istnienia algorytmów wyszukujących tekst. Procesor Intel 8008 zaprezentowany w 1972 posiadał jedynie 14 bitową magistralę adresową co pozwalało na 16 Kbi pamięci. Model Motoroli 68000 posiada 5 MB dysku twardego, co nie może się równać z opecnym standardem darmowej pamięci udostępnianej w chmurze przez Google (15 GB).

Problem wyszukiwania danych nastąpił w momencie tworzenia dużej ilości zawartości. Posiadane archiwum wynosi 14.7 GB danych, niektóre z zawartości są zarchiwizowane co znacznie utrudnia odczytania z nich danych. Nie mniej jednak, posiadane narzędzia w systemach dają dużą dowolność w wyszukiwaniu zawartości, która nas interesuje.

2.2 State of art

Podjęcie problemu wyszukiwania plików po nazwach oraz zawartości jest bardzo złożonym i trudnym problemem w sferze programistycznej. Istnieje wiele rozwiązań tego problemu, które istnieją od początku pracy z komputerem. Narzędzia takie jak `find`, `grep` czy `ffz` [[internet:ToolsRegex](#)] pozwalają na wyszukiwanie zawartości która nas interesuje, ale kompleksowość tych narzędzi nie jest przystosowana do tak trudnego problemu, jakim jest wyszukiwanie treści w plikach, które są zarchiwizowane. Z taką samą niedogodnością spotykamy się w przypadku plików pochodzących z pakietu Microsoft Office 365, jednak jeśli rozwiążemy zadanie otrzymywania zawartości z archiwów, będziemy w stanie otrzymać również zawartość z plików z rozszerzeniami `.doc`, `.docx` czy `.pptx`.

Narzędzie `find` to znane i popularne narzędzie wśród osób zaznajomionych z technologiami linuxowymi. Już bardzo często wykorzystywany do znajdowania plików w systemie, jednak nie nadaje się do znajdowania zawartości plików.

Do przeszukiwania zawartości plików dobrze nadaje się narzędzie `grep`, który jest dostępny w każdej dystrybucji Linuxa. Jego działanie jest dość podobne do `finda`, lecz posiada on możliwość wyszukiwania treści w plikach tekstowych jak również archiwach. Nie posiada on niestety możliwości szukania zawartości plików `.pdf` oraz nie wspiera formatów książkowych takich jak `.djvu`.

Istnieje również `ripgrep`, który jest sukcesorem wcześniej wymienionego narzędzia. Jego wydajność przewyższa `grepa` nawet trzydziestokrotnie w niektórych testach sprawnościowych, jednak zazwyczaj jest to niewielki wzrost. Nie jest on niestety domyślnie instalowany na większości systemów linuxowych. Nie posiada on również wsparcia dla formatów `pdf` i `djvu`.

Można wyszukiwać również po treści piosenek, ale wymagałoby to utworzenie modelu sztucznej inteligencji, która wydobywałaby tekst z piosenek do postaci tekstowej.

2.3 Opis poznanych rozwiązań

2.3.1 Algorytm brute force

Jest wiele algorytmów, które wyszukują tekst. Jednym z takich algorytmów jest algorytm typu `brute-force`. Polega sprawdzaniu każdego bajtu, jego implementacja jest bardzo prosta i standardowa, a złożoność czasowa tego rozwiązania wynosi $O(m * n)$, gdzie m to długość bloku (`pattern`), a n to długość tekstu (`substring`), którego szukamy.

Zaletą tego algorytmu jest to, że nie posiada potrzeby przechowywać żadnych danych w pamięci. Ten algorytm dobrze sprawdza się gdy posiadamy ograniczoną ilość zasobów pamięci, co nie jest problemem w obecnych czasach, gdy pamięć jest stosunkowo tania i szeroko dostępna.

```

1 # Find files by extension:
2     find root_path -name '*.ext'
3
4 # Find files matching multiple path/name patterns:
5     find root_path -path '**/path/**/*.*' -or -name '*pattern*'
6
7 # Find directories matching a given name, in case-insensitive
  mode:
8     find root_path -type d -iname '*lib*'
9
10 # Find files matching a given pattern, excluding specific paths:
11     find root_path -name '*.py' -not -path '*/site-packages/*'
12
13 # Find files matching a given size range, limiting the recursive
  depth to "1":
14     find root_path -maxdepth 1 -size +500k -size -10M
15
16 # Run a command for each file (use '{ }' within the command to
  access the filename):
17     find root_path -name '*.ext' -exec wc -l { } \;
18
19 # Find all files modified today and pass the results to a single
  command as arguments:
20     find root_path -daystart -mtime -1 -exec tar -cvf archive.
      tar { } \+
21
22 # Find empty (0 byte) files and delete them:
23     find root_path -type f -empty -delete

```

Rysunek 2.1: Przykłady użycia programu find

Powyższy algorytm można zrównoleglić, dzieląc wzorzec na mniejsze części i wyszukując tylko dane w tym obszarze, ale należy dołożyć końca wzorca, aby nie wynikła sytuacja, w której wzorzec by wystąpił, ale nie wzięto pod uwagę końca zdania.

Przykład wzorzec ABCABCABDABD podłańcuch BCA rezultat: 4

Jeżeli podzielimy wzorzec na dwa procesy wyszukujące algorytmem brute-force, otrzymamy dwa zadania:

Zadanie 1 Zadanie 2 wzorzec ABCABC wzorzec ABDABD podłańcuch BCA podłańcuch BCA rezultat: -1 rezultat: -1

Zrównoleglenie procesu powoduje, że otrzymaliśmy nie poprawny wynik, gdyż w żadnym z wzorców nie występuje podłańcuch "BCA", choć łańcuch występuje w miejscu 4, to algorytm nie posiada wiedzy o dalszej części wzorca.

Aby poprawić dany algorytm należy dołożyć znaki, które należy sprawdzać w przypadku poprawnego rozpatrzenia ostatniego znaku.

```
1 # Search for a pattern within a file:
2     grep "search_pattern" path/to/file
3
4 # Search for an exact string (disables regular expressions):
5     grep -F|--fixed-strings "exact_string" path/to/file
6 sec
7 # Search for a pattern in all files recursively in a directory,
   showing line numbers of matches, ignoring binary files:
8     grep -r|--recursive -n|--line-number --binary-files without-
   match "search_pattern" path/to/directory
9
10 # Use extended regular expressions (supports '?', '+', '{}', '()'
   ' and '|'), in case-insensitive mode:
11     grep -E|--extended-regexp -i|--ignore-case "search_pattern"
   path/to/file
12
13 # Print 3 lines of context around, before, or after each match:
14     grep --context|before-context|after-context 3 "
   search_pattern" path/to/file
15
16 # Print file name and line number for each match with color
   output:
17     grep -H|--with-filename -n|--line-number --color=always "
   search_pattern" path/to/file
18
19 # Search for lines matching a pattern, printing only the matched
   text:
20     grep -o|--only-matching "search_pattern" path/to/file
21
22 # Search 'stdin' for lines that do not match a pattern:
23     cat path/to/file | grep -v|--invert-match "search_pattern"
```

Rysunek 2.2: Przykłady użycia programu grep

Zadanie 1 Zadanie 2 wzorzec ABCABC(AB) wzorzec ABDABD(nil) podłańcuch BCA
podłańcuch BCA rezultat: 4 rezultat: -1

W takim przypadku sprawdzamy tylko do sytuacji, w której BC jest częścią podłańcu-
cha, ale podłańcuch nie został w pełni znaleziony. Długość ponownego wyszukania byłaby
równa $\text{len}(\text{podłańcuch}) - 1$.

2.3.2 Algorytm Morisa-Pratta

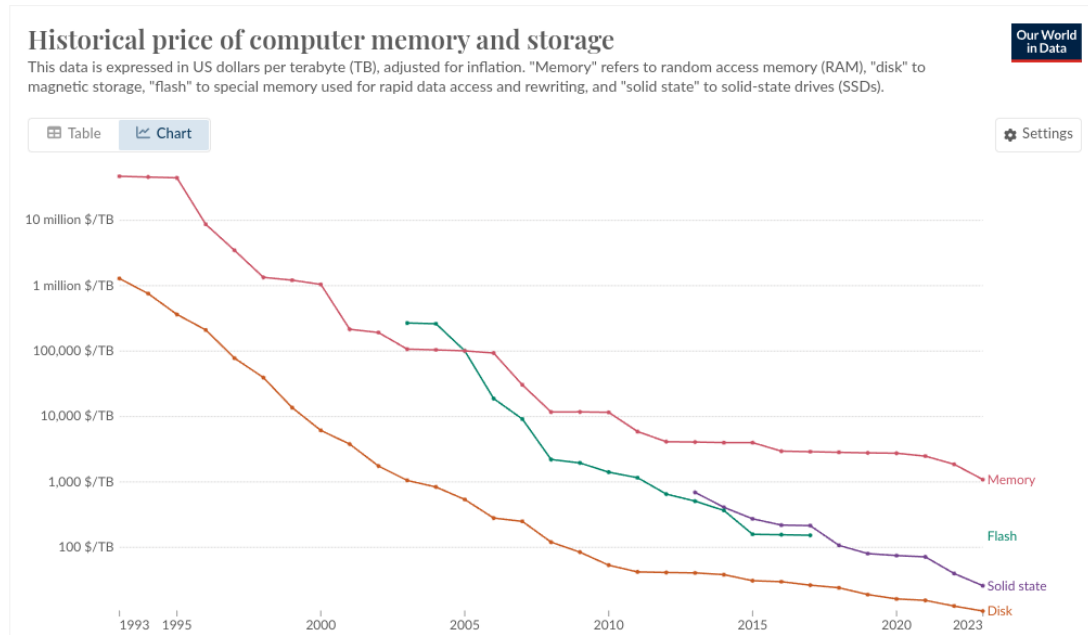
Algorytm Morisa-Pratta jest dość prostym algorytmem wykorzystującym możliwość
wcześniejszego spociesowania podłańcucha wyszukiwanego w tekście co przyspiesza spo-
sób procesowania tekstu. Polega on na wykorzystaniu faktu istnienia pasującego prefikso

```

1 for i := 0; i<len(pattern); i++){
2   for j := 0; j<len(substring); j++){
3     // compare bytes
4   }
5 }

```

Rysunek 2.3: Przykłady algorytmu brute force



Rysunek 2.4: Historyczne dane cen pamięci w latach 1993-2023 [internet:HistoricalMemPrice]

sufiksu. Pozwala to na pominięcie pewnych porównania niektórych znaków, bez szkody w wyniku wyszukiwania.

Dzięki wykorzystaniu tej zależności możemy uniknąć cofania się indeksu i . Tablice `preproc` wypełniamy poprzednią wartością tak długo aż zaistnieje różnica pomiędzy obecnym a następnym znakiem tablicy `substr`. W przypadku różnicy zwiększamy wartość zapisywaną do tablicy preprocesora o odległość różnicy znaków. W ten sposób następnym razem będzie możliwość pominięcia porównania tych znaków.

W drugim etapie można wykorzystać wcześniej przygotowaną tablicę przemieszczeń `preproc`, aby obliczyć ilość przesunięcia w przypadku znalezienia niepasującego prefiksu. Dzięki temu zwykle dłuższy tekst znajdujący się w `s` możemy przeanalizować szybciej niż w przypadku algorytmu brute force. Powoduje to niestety problem w przypadku, gdy napis w którym wyszukujemy nie jest wystarczająco długi.

W podstawowej bibliotece języka Go, w pakiecie `strings` istnieje implementacja metody `Index()`. Nie jest ona jednak w pełni przedstawiona w kodzie, natomiast w jej implementacji można zauważyć, że algorytm brute force jest wykorzystywany tylko w przypadku gdy długość wzorca wynosi więcej niż 64.

```
1 curr = -1
2 preproc[0] = -1
3 for i := 1; i <= len(substr); i++ {
4     for (curr > -1) && (substr[curr] != substr[i-1]) {
5         curr = preproc[curr]
6     }
7     curr++
8     preproc[i] = curr
9 }
```

Rysunek 2.5: Przykład preprocesowania podłańcucha

```
1 res := [] int{}
2 curr := 0
3 found := 0
4 for i := 0; i < len(s); i++ {
5     for (curr > -1) && (substr[curr] != s[i]) {
6         curr = preproc[curr]
7     }
8     curr++
9     if curr == len(substr) {
10        for found < i-curr+1 {
11            found++
12        }
13        res = append(res, found)
14        found++
15        curr = preproc[curr]
16    }
17 }
```

Rysunek 2.6: Przykład preprocesowania podłańcucha

W przypadku w go gdy wzorzec jest większy niż 64 to wykonuje się algorytm podobny do Morisa-Pratta, który jednak posiada dodatkową walidację w przypadku odkrycia false positives. Algorytm Morisa-Pratta nie potrzebuje takiej walidacji.

```
1 func Index(s, substr string) int {
2     switch {
3     case n == 0:
4         return 0
5     [...]
6     case n > len(s):
7         return -1
8     case n <= bytealg.MaxLen: // Zwykle ten case
9         // Use brute force when s and substr both are small
10        if len(s) <= bytealg.MaxBruteForce /* 64 */{
11            return bytealg.IndexString(s, substr)
12        }
13    [...]
14    }
15 }
```

Rysunek 2.7: Szukanie łańcucha w standardowej bibliotece Golang

Rozdział 3

[Tytuł rozdziału]

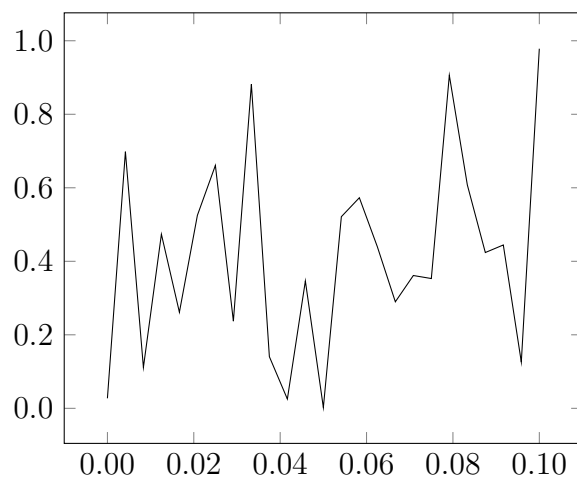
tekst

3.1 [Tytuł podrozdziału]

3.2 [Tytuł podrozdziału]

W całym dokumencie powinny znajdować się odniesienia do zawartych w nim ilustracji (rys. 3.1).

Tekst dokumentu powinien również zawierać odniesienia do tabel (tab. 3.1).



Rysunek 3.1: Wykres przebiegu funkcji.

Tabela 3.1: Opis tabeli nad nią.

ζ	metoda						
	alg. 1	alg. 2	alg. 3			alg. 4, $\gamma = 2$	
			$\alpha = 1.5$	$\alpha = 2$	$\alpha = 3$	$\beta = 0.1$	$\beta = -0.1$
0	8.3250	1.45305	7.5791	14.8517	20.0028	1.16396	1.1365
5	0.6111	2.27126	6.9952	13.8560	18.6064	1.18659	1.1630
10	11.6126	2.69218	6.2520	12.5202	16.8278	1.23180	1.2045
15	0.5665	2.95046	5.7753	11.4588	15.4837	1.25131	1.2614
20	15.8728	3.07225	5.3071	10.3935	13.8738	1.25307	1.2217
25	0.9791	3.19034	5.4575	9.9533	13.0721	1.27104	1.2640
30	2.0228	3.27474	5.7461	9.7164	12.2637	1.33404	1.3209
35	13.4210	3.36086	6.6735	10.0442	12.0270	1.35385	1.3059
40	13.2226	3.36420	7.7248	10.4495	12.0379	1.34919	1.2768
45	12.8445	3.47436	8.5539	10.8552	12.2773	1.42303	1.4362
50	12.9245	3.58228	9.2702	11.2183	12.3990	1.40922	1.3724

Rozdział 4

Badania

Rozdział 5

Podsumowanie

- syntetyczny opis wykonanych prac
- wnioski
- możliwość rozwoju, kontynuacji prac, potencjalne nowe kierunki
- Czy cel pracy zrealizowany?

Bibliografia

- [1] Imię Nazwisko i Imię Nazwisko. *Tytuł strony internetowej*. 2021. URL: <http://gdzies/w/internecie/internet.html> (term. wiz. 30.09.2021).
- [2] Imię Nazwisko, Imię Nazwisko i Imię Nazwisko. „Tytuł artykułu konferencyjnego”. W: *Nazwa konferencji*. 2006, s. 5346–5349.
- [3] Imię Nazwisko, Imię Nazwisko i Imię Nazwisko. „Tytuł artykułu w czasopiśmie”. W: *Tytuł czasopisma* 157.8 (2016), s. 1092–1113.
- [4] Imię Nazwisko, Imię Nazwisko i Imię Nazwisko. *Tytuł książki*. Warszawa: Wydawnictwo, 2017. ISBN: 83-204-3229-9-434.

Dodatki

Dokumentacja techniczna

Spis skrótów i symboli

DNA kwas deoksyrybonukleinowy (ang. *deoxyribonucleic acid*)

MVC model – widok – kontroler (ang. *model-view-controller*)

N liczebność zbioru danych

μ stopnień przyleżności do zbioru

\mathbb{E} zbiór krawędzi grafu

\mathcal{L} transformata Laplace’a

Lista dodatkowych plików, uzupełniających tekst pracy (jeżeli dotyczy)

W systemie do pracy dołączono dodatkowe pliki zawierające:

- źródła programu,
- zbiory danych użyte w eksperymentach,
- film pokazujący działanie opracowanego oprogramowania lub zaprojektowanego i wykonanego urządzenia,
- itp.

Spis rysunków

2.1	Przykłady użycia programu find	5
2.2	Przykłady użycia programu grep	6
2.3	Przykłady użycia programu grep	7
2.4	Historyczne dane cen pamięci w latach 1993-2023 [internet:HistoricalMemPrice]	7
3.1	Wykres przebiegu funkcji.	9

Spis tabel

3.1	Opis tabeli nad nią.	10
-----	------------------------------	----