



**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	[Tytuł rozdziału]	3
3	[Tytuł rozdziału]	7
3.1	[Tytuł podrozdziału]	7
3.2	[Tytuł podrozdziału]	7
4	Badania	9
5	Podsumowanie	11
	Bibliografia	13
	Dokumentacja techniczna	17
	Spis skrótów i symboli	19
	Lista dodatkowych plików, uzupełniających tekst pracy (jeżeli dotyczy)	21
	Spis rysunków	23
	Spis tabel	25

Rozdział 1

Wstęp

Rozdział 2

[Tytuł rozdziału]

Odwołania do literatury: książek [4], artykułów w czasopismach [3], materiałów konferencyjnych [2] i stron www [1].

Równania powinny być numerowane

$$y = \frac{\partial x}{\partial t} \quad (2.1)$$

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 `fdf` 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.

```
1 # Rezultat komendy tldr find
2 # Find files by extension:
3     find root_path -name '*.ext'
4
5 # Find files matching multiple path/name patterns:
6     find root_path -path '**/path/**/*.*ext' -or -name '*pattern*'
7
8 # Find directories matching a given name, in case-insensitive
   mode:
```

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

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`.

```
1 # Rezultat komendy tldr grep
2
3 # Search for a pattern within a file:
4      grep "search_pattern" path/to/file
5
6 # Search for an exact string (disables regular expressions):
7      grep -F|--fixed-strings "exact_string" path/to/file
8
9 # Search for a pattern in all files recursively in a directory,
    showing line numbers of matches, ignoring binary files:
10      grep -r|--recursive -n|--line-number --binary-files without-
        match "search_pattern" path/to/directory
11
```

```

12 # Use extended regular expressions (supports '?', '+', '{}', '()'
    ' and '/' ), in case-insensitive mode:
13     grep -E|--extended-regexp -i|--ignore-case "search_pattern"
        path/to/file
14
15 # Print 3 lines of context around, before, or after each match:
16     grep --context|before-context|after-context 3 "
        search_pattern" path/to/file
17
18 # Print file name and line number for each match with color
    output:
19     grep -H|--with-filename -n|--line-number --color=always "
        search_pattern" path/to/file
20
21 # Search for lines matching a pattern, printing only the matched
    text:
22     grep -o|--only-matching "search_pattern" path/to/file
23
24 # Search 'stdin' for lines that do not match a pattern:
25     cat path/to/file | grep -v|--invert-match "search_pattern"

```

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 posiada on również wsparcia dla formatów pdf i djvu.

Można wyszukiwać również po treści piosenek ale było by to zbyt czasochłonne.

////////////////////////////////////

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.

“code for i := 0; i<len(pattern); i++ for j := 0; j<len(substring); j++ // compare bytes “

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 szerokodostępna. <https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage> <https://jcmit.net/memoryprice.htm>

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.

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 podłańcucha, ale podłańcuch nie został w pełni znaleziony. Długość ponownego wyszukania byłaby równa $\text{len}(\text{podłańcuch}) - 1$.

Rozdział 3

[Tytuł rozdziału]

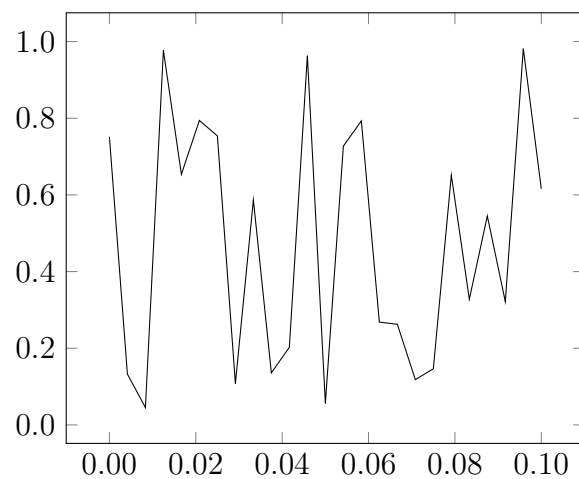
tekst

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

3.1 Wykres przebiegu funkcji.	7
---------------------------------------	---

Spis tabel

3.1	Opis tabeli nad nią.	8
-----	------------------------------	---