Používáte GNU grep? A víte, jak funguje uvnitř?

Ondřej Guth
ondrej.guth@fit.cvut.cz

LinuxDays 2017

1. Úvod
2. Přehled zpracování vstupu
3. Obyčejný řetězec jako regulární výraz
   - BM algoritmus
   - BM a grep
4. Shrnutí

Čím se budeme zabývat

Verze

GNU grep 3.1 (Gentoo GNU/Linux)

Configure

```
--disable-nls --with-included-regex
```

Rychlost

Prohledávání souboru o velikosti 2,1G:

```
time grep kernel /tmp/syslog > /dev/null
real    0m0.023s
user    0m0.001s
sys     0m0.022s
```

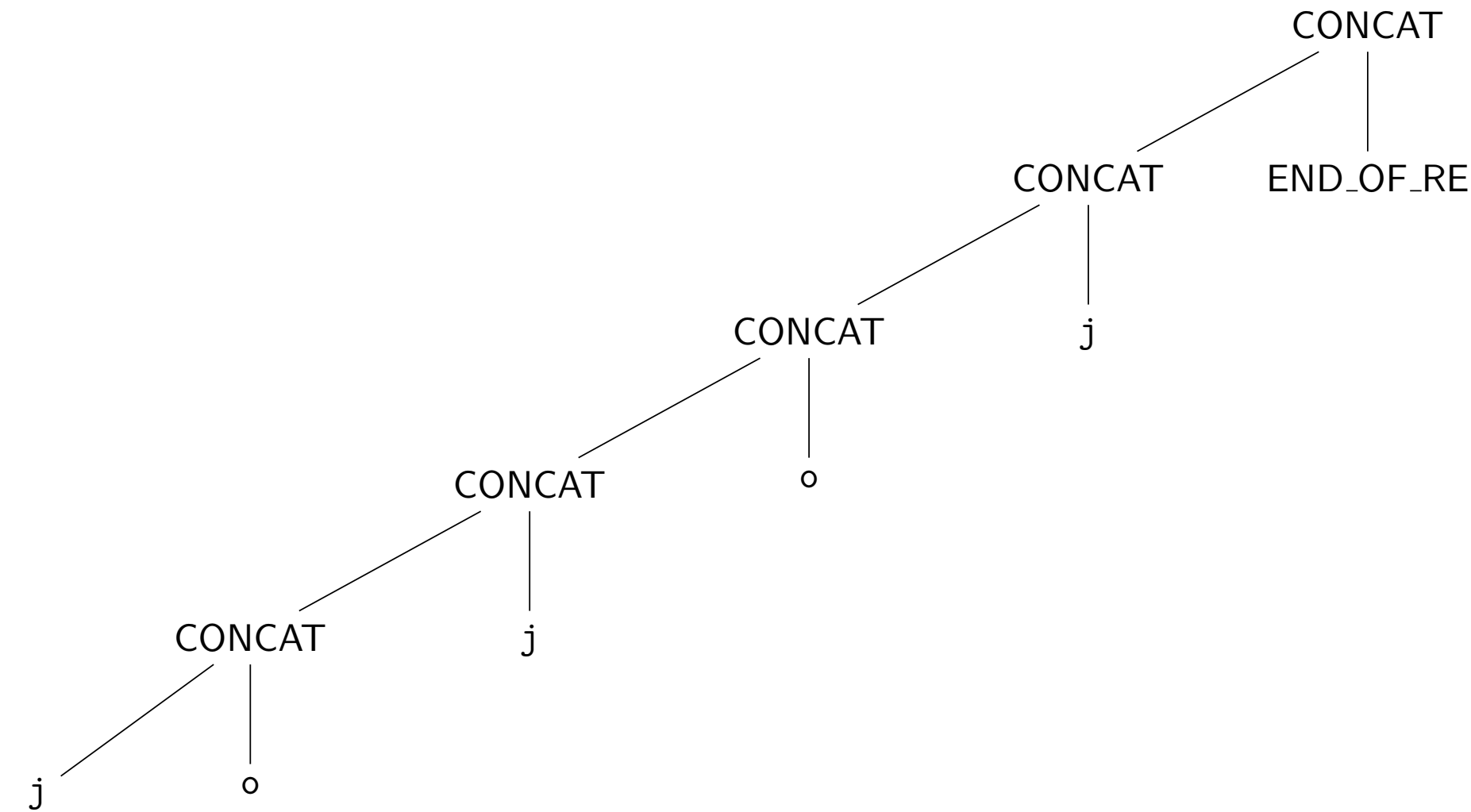
```
time awk /kernel/ /tmp/syslog > /dev/null
real    0m7.233s
user    0m6.850s
sys     0m0.383s
```

Postup

1. Tokenizace výrazu
2. Překlad výrazu do strukturního stromu (libc)
3. Zpracování vstupu prohledávací funkcí
4. V případě nalezeného výskytu vyhledání hranic řádku (a výpis)

Překlad výrazu do strukturního stromu

```
jojoj
```



Algoritmy

Souvislé řetězce v RV

- *kwset matcher*
- jeden vzorek: algoritmus BM
- více vzorků: algoritmus AC

Protisměrné vyhledávání

Méně porovnání než je znaků vstupu.

Příklad

```
j o j k j o j o j
j o j o j
        j o j o j
```

Algoritmus Boyer-Moore

Varianta použitá v GNU grep kwset matcher

Posun v případě neshody:

Bad character shift

Nejbližší výskyt prvního přečteného znaku (zprava) z textu ve výrazu.

delta2

Zarovnání výrazu na další opakování dosud přečtené části textu.

Algoritmus Boyer-Moore

Bad character shift

BCS pro `jojoj`

`j`: 0
`o`: 1
`k`: 5

Příklad

```
j o j o o k j o j o j
j o j o j   j o j o j
  j o j o j
```

Posun: 1. Posun: 5. Posun: 0.

Algoritmus Boyer-Moore

Varianta použitá v GNU grep kwset matcher

Posun v případě neshody:

Bad character shift

Nejbližší výskyt prvního přečteného znaku (zprava) z textu ve výrazu.

delta2

Zarovnání výrazu na další opakování dosud přečtené části textu.

Algoritmus Boyer-Moore

delta2

Příklad

K	Y	K	Y	R	Y	K	Y
0	6	6	6	6	4	6	2

Zpracování vstupu v GNU grep

Příklad

```
K Y K Y K Y K Y R Y K Y\nY K Y K Y R Y K Y K Y R Y K Y K Y
K Y K Y K Y K Y R Y K Y   K K K K K K K K Y
```

BCS pro `Y` je 0 delta2 pro `YKY` je 4 výkyt, hranice řádku (memchr a memrchr), výpis, další řádek BCS pro `K` je 1

delta2

```
K Y K Y R Y K Y
0 6 6 6 6 4 6 2
```

Závěr

- hledání s kwset matcher pro jeden vzorek – BM algoritmus
- hledání s kwset matcher pro více vzorků (grep -e vz1 -e vz2) – AC algoritmus
- počítání tabulek pro posuvy
- komplikovanější vzorky