Accelerating Knuth-Morris-Pratt String matching over LZ77 Compressed Text

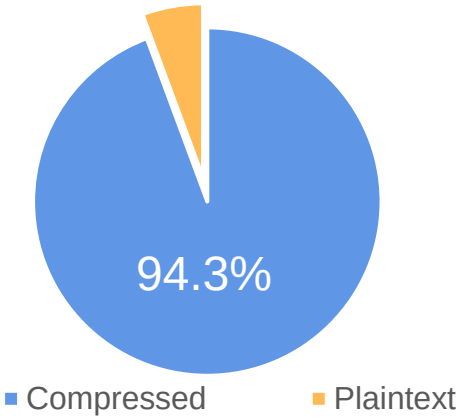
Xiuwen Sun, Di Wu, Da Mo, Jie Cui, Hong Zhong

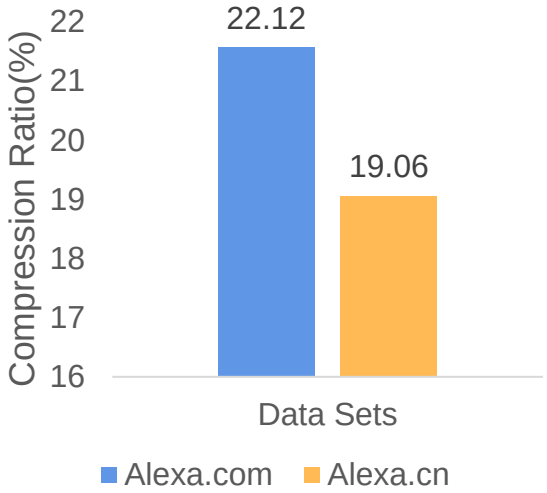
Anhui University

Background

✓**90%**: more than 90% web sites are compressed.

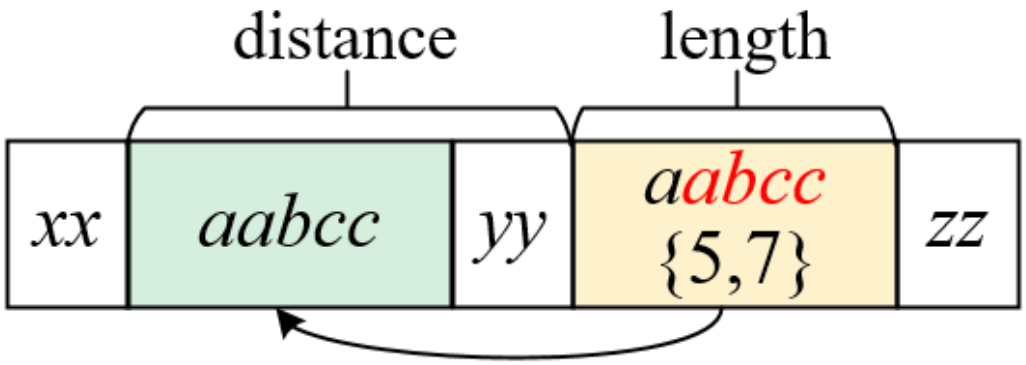
✓**20%**: the average compression ratio of web pages.





Design Principle

- ✓ LZ77 replaces redundancy by a two-tuple, {***length, distance***}.
- ✓ The two-tuple is called **pointer**;
- ✓ The corresponding substring is called **refferd string**.
- ✓ More than 91% bytes are represented within pointers witch can be skipped.



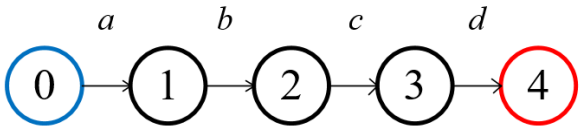
Basic Idea

- **auxiliary function *σ(x)***: the length of the longest prefix of pattern *P* that is also a suffix of string *x*.

$P = abcd$

$\sigma(aa) = 1,\ \sigma(aabc) = 3$

- **A** = (Q, Σ, δ, , F): string matching automaton that corresponds to *P*[1....m].

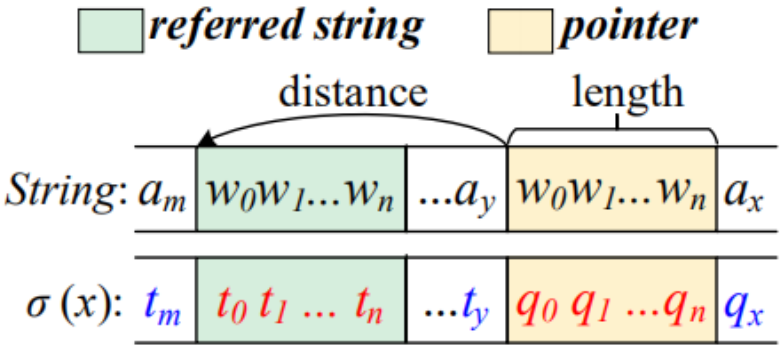


$Q = \{0,1,\cdots,m\}$, $q_0 = 0$, $F = \{m\}$, $\delta(q,a) = \sigma(P_q a)$ $q \in Q$, $P_q = P[1...q]$

⇨ For any string ***x*** and character ***a***,

if $q = \sigma(x)$, then

Algorithm



$t_y = t_m$

copy to

if q_i is accepting state, matching pattern

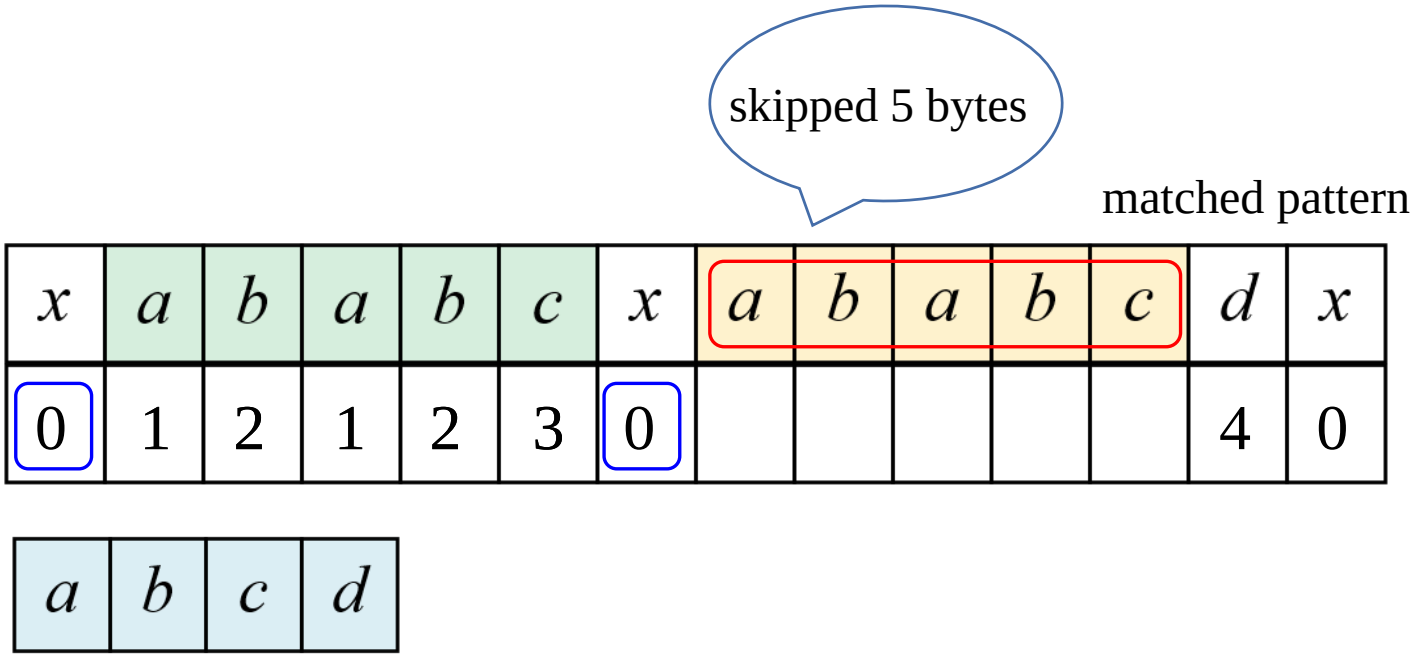
$t_m \neq t_y$

scan w_0 in *pointer*, assume get $q_0 = t_0$

$q_1 = \sigma(\mathrm{P}_{q_0}\mathrm{w}_1) = \sigma(\mathrm{P}_{t_0}\mathrm{w}_1) = \mathrm{t}_1$

if q_i is accepting state, matching pattern

Example



Evaluation Settings

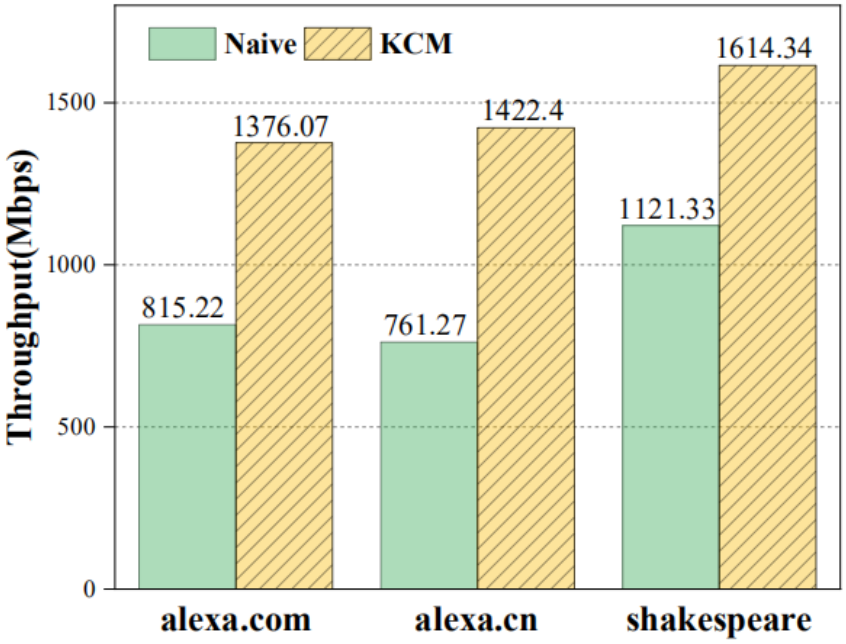
Platform: Intel i5-9400 CPU 2.9 GHz, 16 GB RAM

Rule Set: Snort rules, Roles'names of Shakespeare

Data Set: Homepages of Alexa top sites, Shakespeare collected works

Characteristic	Alexa.com	Alexa.cn	Shakespeare
Count of Compressed Pages	434	13747	11
Compressed Size (MB)	15.54	226.95	0.55
Decompressed Size (MB)	70.24	1190.99	1.44
Bytes represented by pointers	**91.21%**	**91.92%**	**94.06%**
Average pointer length (B)	14.89	19.84	6.37

Performance



Throughput of the methods over three data sets

Characteristic	Alexa.com	Alexa.cn	Shakespeare
matched patterns	155,087	2,439,445	4976
skipped ratio	91.07%	91.85%	93.99%
Bytes represented by pointers	**91.21%**	**91.92%**	**94.06%**
throughput boost	1.69x	1.87x	1.44x

Metrics of KCM over three data sets

Conclusion

KCM

A method for KMP string matching over compressed traffic.

The skipped bytes approaches the theoretical upper bound.

Faster

1.58 Gbps on matching compressed text, which is 1.4~1.9 times improvement under the experiments with real traffic.