On Average Sequence Complexity

Svante Janson
Uppsala University, Sweden
Stefano Lonardi
University of California, Riverside, CA
Wojciech Szpankowski
Purdue University, West Lafayette, IN

Subword complexity

- The *subword complexity* ***c(x)*** of a string ***x*** is the number of distinct substrings of ***x*** (of any length)
- Also called *"complexity index"* or "*linguistic complexity*"
- Intuition: sequence with low complexity tend to be "repetitive" (easy to compress)

Subword complexity

sliding window

text

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Subword complexity

sliding window ***x***

text

- Build the suffix tree ***T*** on ***x***
- Compute the subword complexity ***c(x)*** by counting #implicit+#explicit-1 nodes in ***T***

Is ***c(x)*** statistically significant?

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Probabilistic model

- To assess the statistical significance, we need a probabilistic model for the source
- We assume that a Markov model is generating the text
- We associate a random variable to ***c(x)***

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Random variable $C_{n,k}$

- Let $\boldsymbol{C_{n,k}}$ be the random variable associated to the subword complexity of a string of size ***n*** over an alphabet of cardinality ***k***
- We would like to compute

$$\boxed{z(x) = \frac{c(x) - E(C_{n,k})}{\sqrt{Var(C_{n,k})}}}$$

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Objective

- Goal: characterize precisely ***E***$(C_{n,k})$
- The characterization of ***Var***$(C_{n,k})$ is left as an exercise to the audience ☺

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Maximum complexity

- It is easy to realize that

$$\boxed{c(x) \leq \sum_{l=1}^{n} \min(k^l, n-l+1)}$$

- When $n=k$

$$\boxed{c(x) \leq \frac{n(n+1)}{2}}$$

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Related works

- In 1993, J. Shallit [Graphs and Comb.] proved that for ***k=2***

$$\boxed{c(x) \le \frac{(n-d+1)(n-d)}{2} + 2^{d+1} - 1 \approx \frac{n^2}{2}}$$

where ***d*** is the unique integer s.t.

$$\boxed{2^d + d - 1 \le n \le 2^{d+1} + d}$$

- Upper bound can be attained (de Brujin graph)
- Extension to larger alphabets [FHS04]

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Related works

- In 1999, A. de Luca [TCS] proved that

$$\boxed{c(x) = 1 + \frac{(n+K)(n-K+1)}{2} - \sum_{j=2}^{k}\sum_{i=0}^{R} ig(j,i)}$$

where ***K*** is the length of the shortest suffix that occurs only once, ***H*** is the length of the shortest prefix that occurs only once, ***R-1*** is the height of the deepest branching node in the suffix tree for ***x***, ***L-1*** is the height of the deepest branching node in the suffix tree for x^R, and ***g(j,i)*** is the count of the words of length I which are branching nodes with at least j children

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Related works

- Kása [Pure Math. Appl.] studied $C_{k,k}$ for short strings and conjectured that

If one chooses $k = \frac{l(l+1)}{2} + 2 + i$ where $l \geq 2$ and $0 \leq i \leq l$, then

$P(C_{k,k} = t) > 0$ for all t such that $\frac{l(l^2-1)}{2} + 3l + 2 + i(l+1) \leq t \leq \frac{k(k+1)}{2}$

- Proved by Leve and Séébold [Bull. Belgian Math. Soc.]

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Example *n=4, k=2*

$C_{4,2}$	*f*	*words*
4	2	AAAA BBBB
5	0	
6	0	
7	6	AAAB ABAB ABBB BAAA BABA BBBA
8	8	AABA AABB ABAA ABBA BAAB BABB BBAA BBAB

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Example *n=4, k=3*

$C_{4,3}$	*f*	*words*
4	3	AAAA BBBB CCCC
5	0	
6	0	
7	18	AAAB AAAC ABAB ABBB ACAC ACCC BAAA BABA BBBA BBBC BCBC BCCC CAAA CACA CBBB CBCB CCCA CCCB
8	24	AABA AABB AACA AACC ABAA ABBA ACAA ACCA BAAB BABB BBAA BBAB BBCB BBCC BCBB BCCB CAAC CACC CBBC CBCC CCAA CCAC CCBB CCBC
9	36	AABC AACB ABAC ABBC ABCA ABCB ABCC ACAB ACBA ACBB ACBC ACCB BAAC BABC BACA BACB BACC BBAC BBCA BCAA BCAB BCAC BCBA BCCA CAAB CABA CABB CABC CACB CBAA CBAB CBAC CBBA CBCA CCAB CCBA

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Summary for *n=4, k=2...10*

	$C_{4,2}$	$C_{4,3}$	$C_{4,4}$	$C_{4,5}$	$C_{4,6}$	$C_{4,7}$	$C_{4,8}$	$C_{4,9}$	$C_{4,10}$
4	2	3	4	5	6	7	8	9	10
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	6	18	36	60	90	126	168	216	260
8	8	24	48	80	120	168	224	288	370
9		36	144	360	720	1260	2016	3024	4320
10			24	120	360	840	1680	3024	5040

$$c(x) \le \min(k,4) + \min(k^2,3) + \min(k^3,2) + \min(k^4,1)$$

when $k \ge 4$, then $c(x) \le 4+3+2+1=10$

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Exact distribution for *n<6*, any *k*

$n \to$	2	3	4	5
$\mathbf{P}(C_{n,k} = 2)$	$1/k$	0	0	0
$\mathbf{P}(C_{n,k} = 3)$	$1-1/k$	$1/k^2$	0	0
$\mathbf{P}(C_{n,k} = 4)$	0	0	$1/k^3$	0
$\mathbf{P}(C_{n,k} = 5)$	0	$3(k-1)/k^2$	0	$1/k^4$
$\mathbf{P}(C_{n,k} = 6)$	0	$(k-1)(k-2)/k^2$	0	0
$\mathbf{P}(C_{n,k} = 7)$	0	0	$3(k-1)/k^3$	0
$\mathbf{P}(C_{n,k} = 8)$	0	0	$4(k-1)/k^3$	0
$\mathbf{P}(C_{n,k} = 9)$	0	0	$6(k-1)(k-2)/k^3$	$3(k-1)/k^4$
$\mathbf{P}(C_{n,k} = 10)$	0	0	$(k-1)(k-2)(k-3)/k^3$	0
$\mathbf{P}(C_{n,k} = 11)$	0	0	0	$10(k-1)/k^4$
$\mathbf{P}(C_{n,k} = 12)$	0	0	0	$2(k-1)(3k-5)/k^4$
$\mathbf{P}(C_{n,k} = 13)$	0	0	0	$19(k-1)(k-2)/k^4$
$\mathbf{P}(C_{n,k} = 14)$	0	0	0	$10(k-1)(k-2)(k-3)/k^4$
$\mathbf{P}(C_{n,k} = 15)$	0	0	0	$(k-1)(k-2)(k-3)(k-4)/k^4$
$\mathbf{P}(C_{n,k} \geq 16)$	0	0	0	0

? Expectation and Variance (assumes uniform i.i.d.)

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Main results (1/2)

Theorem 1. *Let $C_{n,k}$ be the complexity index of a string of length n generated by a strongly mixing stationary source. Then, for large n,*

$$\mathbf{E}C_{n,k} = \binom{n+1}{2} - O(n \log n).$$

Hence $C_{n,k} = n^2/2 + O_p(n \log n)$, i.e. $(n^2/2 - C_{n,k})/n \log n$ is bounded in probability.

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Proof sketch (1/4)

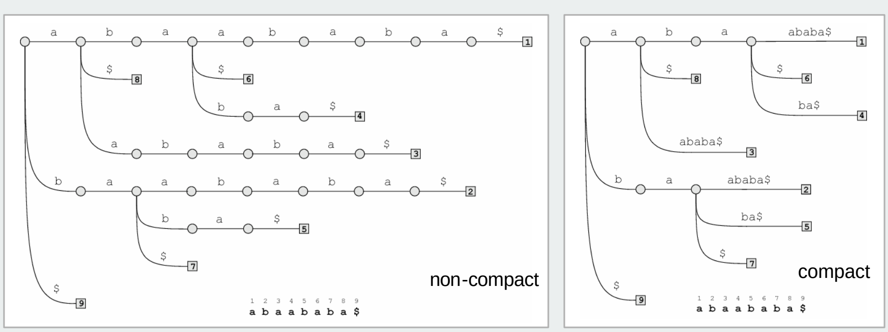
- Want to prove that

  $n(n+1)/2-\mathbf{E}(\mathbf{C}_{n,k}) \in O(n \log n)$

- Consider the non-compact trie and the compact trie for the same string

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Proof sketch (2/4)



- Recall that $c(x)$=#implicit+#explicit nodes-1

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Proof sketch (3/4)

- The quantity $\boldsymbol{C_{n,k}}$ is mostly determined by the nodes in the non-compact tries that are not in the compact trie (at the bottom)
- Consider the *i*-th suffix

Nodes which are not in the compact trie are n-i-H_n where H_n is the *height* of the trie

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Proof sketch (4/4)

We have $C_{n,k} \geq \sum_{i=0}^{n-1} (n - i - H_n) = \frac{n(n+1)}{2} - nH_n$.

Then $\frac{n(n+1)}{2} - E(C_{n,k}) \leq nE(H_n)$. The quantity $E(H_n)$ is well-studied (see, e.g., Spza [SICOMP'93]) and is known to be $O(\log n)$.

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Main results (2/2)

Theorem 2. *Let $C_{n,k}$ be the complexity index of a string generated by an unbiased memoryless source. Then the average l-subword complexity is*

$$\mathbf{E}(C^l_{n,k}) = k^l(1 - e^{-nk^{-l}}) + O(l) + O(nlk^{-l}).$$

Furthermore, for large n the average complexity index becomes

$$\mathbf{E}(C_{n,k}) = \binom{n+1}{2} - n\log_k n + \left(\frac{1}{2} + \frac{1-\gamma}{\ln k} + \phi_k(\log_k n)\right)n + O(\sqrt{n\log n})$$

where $\gamma \approx 0.577$ is Euler's constant and

$$\phi_k(x) = -\frac{1}{\ln k}\sum_{j\neq 0}\Gamma\left(-1 - \frac{2\pi i j}{\ln k}\right)e^{2\pi i j x}$$

is a continuous function with period 1. $|\phi_k(x)|$ is very small for small k: $|\phi_2(x)| < 2\cdot 10^{-7}$, $|\phi_3(x)| < 5\cdot 10^{-5}$, $|\phi_4(x)| < 3\cdot 10^{-4}$.

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside

Findings

- Closed form for probability distribution of $\boldsymbol{C_{n,k}}$ for $n<6$, any k
- Asymptotic results for strongly mixing stationary source
- More accurate asymptotic results for unbiased memory-less source

Stefano Lonardi
Department of CS and E
Bourns College of Engineering
University of California, Riverside