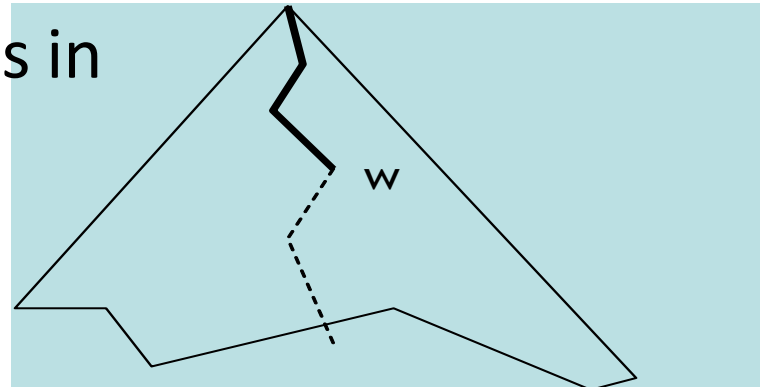
Average Size of a Suffix Tree for Markov Sources

Philippe Jacquet (Nokia Bell Labs)

Wojciech Szpankowski (Purdue U.)

Tries and Suffix tree

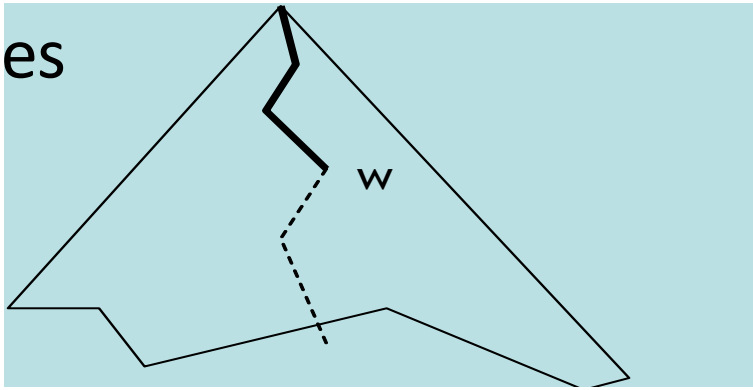
- Trees are fundamental structures in computer science
  - n binary sequences: X_1, X_2,…, X_n
  - Tries(X_1,…,X_n) is a binary tree
    - $w \in \{0,1\}^*$ is a path to a node in Tries(X1,…,Xn)
    - If w is prefix of at least two sequence Xi and Xj.
    - Used in fast retrieval in IP routing table
  - Obvious generalization for any finite alphabet $\mathfrak{A}$
    - $|\mathfrak{A}|$-ary tree



$X_i = w\cdots, X_j = w\cdots$

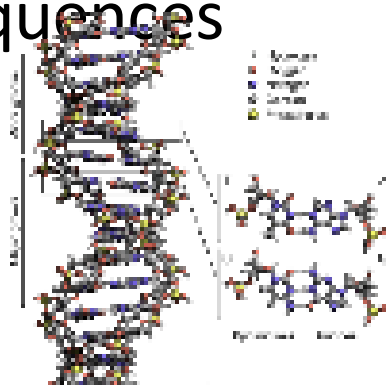
Suffix trees

- Suffix Tree is a Special case of Tries
  - X sequence of length n
  - n first suffixes $S_1, S_2, \ldots, S_n$
  - SuffixTree(X,n)=Tries($S_1,\ldots,S_n$)
  - $w \in \mathfrak{A}^*$ is a node in the SuffixTree(X)
  - If w is at least a double factor in X



$S_i = w\cdots, S_j = w\cdots$

- Suffix tree is used Ziv-Lempel 77 compression
- Used for pattern matching between DNA sequences
  - $\mathfrak{A}$ ={G,A,T,C}



Source models for Tries and suffix trees

- Uniform Binary memoryless Bernoulli for tries
  - Adapted for IP routing

Flajolet, P., & Sedgewick, R. (1986). Digital search trees revisited. *SIAM Journal on Computing,*.

- General memoryless bernoulli on finite alphabet for suffix tree

JS94 Jacquet, P., & Szpankowski, W. (1994). Autocorrelation on words and its applications: analysis of suffix trees by string-ruler approach. *Journal of Combinatorial Theory*.

  - Uniform 4-ary adapted for DNA sequencing
- General Markov on finite alphabet for suffix trees
  - Adapted for suffix tree in natural text compression

Markov model of natural text

- $\mathfrak{A}$={a,b,c,d,e,…}
- Memoryless text generation
  - esdehTe,a; psseCed vcenseusirh vra f uetaiapgnuev n cosb mgffgfL itbahhr nijue n S ueef,ru s,k smodpztrnno.eeteespfg mtet tr i aur oiyr
- Markov text generation (memory 3)
  - We hat Government of Governments long that their right of abuses are these rights, it, and or themselves and are disposed according Men, der.

Le reconnaissez-vous ?

Memory 0

RAAS IEILC' 0 AEU
AOS
TLD ARFPLRTEE FAEP
V SUEHELLESNEA'OATC STTQAA
EAQEEUZEPTOUETALEFIEHFUSEZELPEJTOIPIRNGEE
T
HLTE OE' STSLENE
LTAE U

NAPEDSEBI SESESN OTO REMDOUUOFJANIRP
C HHE AGAINEALJE EPFPPIFSSAIU EIEMTSORI
EEUESTAOGD A NUELLQEEIT'T REQVEH SNENE TS
ACOSSEU NAESDUAV FRI
S TEN MLPN LETEOUE
D N DTSMAAE C L ER TEYSAT TUNA SZFTPAZ
OLMAENPARTLOTUARU VHDZA
UUENR CLI
T D C SSETVAD Q NLL
DAUTTNOIO FPVIERSLF NE UCRAT
IEREEEOLSNIUE"IDITTPLEAIRSUTNOES PEAS
AT AAHVR
L SA E IVAVEDUEASMEFM
ETVSNT ELIUNTUUR EI SUSQV ASTSTTLUEN S
JANIETNN TSSEE TT
SS OCI
ASJAEN

Memory 1

LQUSERI FOU' A FAIS VE LLL FONT L'OINERA LET PAN
SOUT LLETITE SURE
SELQURISU PRMOITRT VOUE CHASTA LET CISA CISE D
VOUR
CE
S CHA MOUS MONERE
CHAVUBI C''A CHEANT DEL URT TORITIS VOR VEZ GASI
PRT TESE
CRTAN MOUER PLLE
DEZ PRONA DIE VAN LLLQUS PRCHANINTE
N SUEPA CHA VALETISA DE
LANIGRCHEMA DET FAIANS VUE FOINDE LER
QUE LEPRAUE PAl PA ONE VOUD
QUVOUBIAI,
QUTESEUA FAI VOURIENTENDANIPLEPA PR
PAMI DI STIS DE E
VOR
QUN'E BIPE
VORINOUS PR DISERMONER VELUT
E
CHAVE
QUBISEMAIE
D'A AL'E AINET VUS DERTE
E JEMISET PR TTE DA E PAVU JOUSETINALE T
OUSELLET LLANOURVOUIE
CRMPANSOUBITE LEN LA

Memory 2

LA SUIT ETTERAISISSE
AVANT PRETIER SUI DANTEUS CHE ALLE OURMINE
JE FAMI DEPLA VER
JE DANT DE AIER
QUE
QUE
LAI D'A SE
LA FAINE
ELLA PAS ALE AYANSEAU DE FOINCIGALLE
EMPRIANTER AYANT LA SA TOURVUE
PAS CIPAS PRIND LA PAIND L'EST ELLA CIGALE FUT
ELLA FUT
QUE
LA CHANTE
VOURVUE GRAI SUBSIN MOI PRE DANTIT DEFAMINT
QU'A LUI N'EST MORCEAU
ELLE
C'EST CHANSEUSE
PAS A CHANSE
VOUSEAU DIT VENUELLE VEN PETE
AVA L'OURMI PRETENUI PRET J'EN PRIER
QUE FOUR SAISE CHANT POURVUELLE LA CIPALE
CHANT ETEMPS UN SA TENUIS
DE AIER FOUSE VOUVEI.LE VENANT EMPS AU DANTE
QUE AL
LA SA I.A FORT DEPLAIER

Memory 3

LA PRET PAS PAS UN SUBSISTERETE
TOUT VENANT
LA SAISON MORCEAU
ELLE AYANT
LA SON MOINDRE DEPOURMI N'EST L'OUT VENUE
PAS UN SUIS NE VOISINE
CHEZ J'EN DANSEZ MAIN POURMI SAIS NE VOISIEZ VOISINE
LA FORT AISINE
CHEZ LA CIGALE A TOUT VERMI N'EST L'OUT L'ETER
QUELQUE GRAINTERETEUSE
VOUS DEPOUR SUIS FOI D'ANIMAL
INTE
TOUT LA CIGALE AYANTE
TOUT LA CIGALE ALLA SA VOUS PRIANT
JE CHANT CHAUD
DIT ET JOURMI N'EST L'ETEUSE
NUIT ET PAIERAI LUI D'ANIMAL
INTEUSE
EH BIEN SUBSISTERETE
SE FAMINE
LA SA VOISIEZ VOISINE
CHE OU DEFAUT
QUELQUE GRAI UJI PRIANTIEZ J'EN SEUL PETIT ET ET PRETE
SE

Memory 4

LA CIGALE AYANT L'OUT FOI D'ANIMAL
INTENANT
LA CRIER FAMINE
LA CIGALE A CETTE EMPRUNTEUSE
NUIT ELLE
JE VOUS PAIERAI LUI DIT ET PRIANT DE VERMISSEAU
ELLE
JE VOUS DEPLAISE
VOUS CHANTAIS NE VOUS DEPOURVUE
QUAND LA BISE FUT VENANT
JE CHAUD
DIT ELLE
AVANT DEPOURVUE
QUAND LA FOURMI SA VOISINE
LA PRIANT L'ETE
SE TROUVA FOURMI N'EST LA SAISON MOINDRE
DEFAUT
QUE FAISIEZ VOUS AU TEMPS CHAUD
DIT ET PRIANT CHANTE
TOUT FOI D'ANIMAL
INTENANT
LA CRIER FAMINE
CHEZ LA BISE FUT VENUE
PAS UN SEUL PETIT MORCEAU
DE MOUCHE OU DE VERMISSEAU
ELLE AYANT DEPLAISE
VOUS DEPLAISE
VOUS PAIERAI LU

Our result

- t_n average size of tries
  - of n independent Markov sequences
- s_n average size of suffix tree
  - on n first suffixes of a Markov sequence
- Convergence $t_n - s_n = O(n^{1-\varepsilon})$
- We have $t_n = \frac{n}{h} + o(n)$ in general
  - h per symbol entropy rate of the Markov sequence
  - $t_n = \frac{n}{h} + nP_2(\log n) + O(n^{1-\varepsilon'})$ in some periodic case

Markov models on digital trees

- Tries under dynamic sources

Clément, J., Flajolet, P., & Vallée, B. (2001). Dynamical sources in information theory: A general analysis of trie structures. *Algorithmica*,.

- Markovian Digital search trees

Jacquet, P., & Szpankowski, W. (1991). Analysis of digital tries with Markovian dependency. *IEEE Transactions on Information Theory*.

- Insertion Depth in Markov suffix trees

FW05 Fayolle, J., & Ward, M. D. (2005). Analysis of the average depth in a suffix tree under a Markov model. In *International Conference on Analysis of Algorithms DMTCS*

Analysis

- $t_n = \sum_{w \in \mathfrak{A}^*} P(N_n(w) \geq 2)$
  - $N_n(w)$ is the number of sequences prefixed by w
- $s_n = \sum_{w \in \mathfrak{A}^*} P(O_n(w) \geq 2)$
  - $O_n(w)$ is the number of positions before n where w appears

$$P(N_n(w) \geq 2) = 1 - (1 - P(w))^n - nP(w)(1 - P(w))^{n-1}$$

$P(N_n(w) \geq 2)$ is much more complicated…

Analysis

- $t_n - s_n = \sum_{w \in \mathfrak{A}^*} q_n(w) + d_n(w)$
  - With $q_n(w) = P(O_n(w) = 1) - nP(w)(1 - P(w))^{n-1}$
  - And $d_n(w) = P(O_n(w) = 0) - (1 - P(w))^n$
  - The term $\sum_{w \in \mathfrak{A}^*} q_n(w)$ is already studied in FW05
    - Indeed the average depth

$$E\left[D_n^S\right] - E\left[D_n^T\right] = \frac{1}{n}\left(\sum_{w \in \mathfrak{A}^*} P(O_n(w) = 1) - P(N_n(w) = 1)\right) = \frac{1}{n}\sum_{w \in \mathfrak{A}^*} q_n(w)$$

    - And is proven to be $O(n^{-\varepsilon})$ as in JS94 for the memoryless case.

Analysis

- Via generating function

$$Q(z) = \sum_{n, w \in \mathfrak{A}^*} q_n(w) z^n = \sum_{w \in \mathfrak{A}^*} P(w) \left(\frac{z^{|w|}}{D_w(z)^2} - \frac{z}{\left(1-(1-P(w))z\right)^2} \right)$$

  - $S_w(z)$ the autocorrelation polynomial of word w.
  - $D_w(z) = S_w(z)(1-z) + z^{|w|}\left(1+(1-z)F_w(z)\right)$
  - With $w = bua : F_w(z) = \frac{1}{\pi_a}\left[(\mathbf{P} - \boldsymbol{\pi} \otimes \mathbf{1})(\mathbf{I} - z(\mathbf{P} + \boldsymbol{\pi} \otimes \mathbf{1}))^{-1}\right]_{b,a}$
    - Term $F_w(z)$ comes from the analysis in

Régnier, M., & Szpankowski, W. (1998). On pattern frequency occurrences in a Markovian sequence. *Algorithmica*.

    - In memoryless case $F_w(z)=0$ and is done in JS94

Analysis

- $$\sum_{w\in\mathfrak{A}^*} q_n(w) = \frac{1}{2i\pi}\oint Q(z)\frac{dz}{z^{n+1}} = \mathrm{res}(Q(z), A_w)A_w^{-n} + O(\rho^{-n})$$
  - A_w first root of $D_w(z)$ is in $\delta^{|w|}$
  - Conclusion $\sum_{w\in\mathfrak{A}^*} q_n(w) = O(n^{1-\varepsilon})$ comes as in JS94
    - because Fw(z) is uniformly bounded when z in any compact set such that $|z| < |\lambda|$, second eigenvalue of $\mathbf{P}$

Analysis

- The case

$$\Delta_w(z) = \sum_n d_n(w) z^n = \frac{P(w)z}{(1-z)} \left(\frac{1+(1-z)F_w(z)}{D_w(z)} - \frac{1}{\left(1-(1-P(w))z\right)} \right)$$

  - poses problem because

$$\sum_{w \in \mathfrak{A}^*} q_n(w) = \sum_{w \in \mathfrak{A}^*} \mathrm{res}(\Delta_w(z), A_w) - (1-P(w))^n + O(\rho^{-n})$$

  - Leads to a term $\sum_{w \in \mathfrak{A}^*} P(w) \frac{F_w(A_w)}{D_w(A_w)}$ which might be infinite (bounding $F_w(A_w)$ as in FW05 is not sufficient)

Analysis (end)

- But $\sum_{w\in\mathfrak{A}^{k+1}} P(w)F_w(z) = \text{trace}(\mathbf{M}(z)\mathbf{P}^k)$
  - With $\mathbf{M}(z) = (\mathbf{P} - \boldsymbol{\pi}\otimes\mathbf{1})(\mathbf{I} - z(\mathbf{P} + \boldsymbol{\pi}\otimes\mathbf{1}))^{-1}$
  - Since $\mathbf{P}^k = \boldsymbol{\pi}\otimes\mathbf{1} + O(\lambda^k)$ and $\boldsymbol{\pi}\otimes\mathbf{1}\mathbf{M}(z) = \mathbf{M}(z)\boldsymbol{\pi}\otimes\mathbf{1} = 0$

$$\text{trace}(\mathbf{M}(z)\mathbf{P}^k) = O(\lambda^k)$$

  - The term $\sum_{w\in\mathfrak{A}^*} P(w)\frac{F_w(A_w)}{D_w(A_w)} = \sum_k \sum_{w\in\mathfrak{A}^{k+1}} P(w)F_w(1) + O(\delta^k)$ is bounded
  - We got the equivalence suffix tree/ tries in Markovian source model

Tries under Markovian model

- The only truly new result is the characterization of asymptotic periodic case.
  - We have $t_n = \frac{n}{h} + nP_2(\log n) + O(n^{1-\varepsilon'})$ with P2 periodic when
  - $\forall (a,b,c) \in \mathfrak{A}^3 : \log\left(\frac{P_{b,a}P_{a,c}}{P_{b,c}}\right)$ are commensurable.