Referências Bibliográficas

[Adl01] ADLER, M.; MITZENMACHER, M.. **Towards compressing web graphs**. In: DATA COMPRESSION CONFERENCE, 2001. PROCEEDINGS. DCC 2001., p. 203–212, 2001. 4

[Alb99] ALBERT, R.; JEONG, H. ; BARABASI, A.-L.. **Internet: Diameter of the world-wide web**. Nature, 401(6749):130–131, September 1999. 4

[Asa08] ASANO, Y.; MIYAWAKI, Y. ; NISHIZEKI, T.. **Efficient compression of web graphs**. p. 1–11. 2008. 4

[Avi09] AVILA, B. T.; LABER, E. S.. **Merge source coding**. In: ISIT'09: PROCEEDINGS OF IEEE INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY, p. 1–5, Seoul, South Korea, 2009. 5

[Bay72] BAYER, R.; MCCREIGHT, E.. **Organization and maintenance of large ordered indexes**. Acta Informatica, 3(1):173–189, 1972. 4.2

[Bay77] BAYER, R.; UNTERAUER, K.. **Prefix b-trees**. ACM Trans. Database Syst., 2:11–26, March 1977. 4.1

[Ben76] BENTLEY, J. L.; YAO, A. C. C.. **An almost optimal algorithm for unbounded searching**. Information processing letters, 5(3):82–87, 1976. 3

[Ben05] BENDER, M. A.; DEMAINE, E. D. ; FARACH-COLTON, M.. **Cache-oblivious b-trees**. SIAM Journal on Computing, 35(2):341, 2005. 4.2, 4.2.3

[Bha98] BHARAT, K.; BRODER, A.; HENZINGER, M.; KUMAR, P. ; VENKATASUBRAMANIAN, S.. **The connectivity server: fast access to linkage information on the web**. Comput. Netw. ISDN Syst., 30(1-7):469–477, 1998. 4

[Bol04a] BOLDI, P.; VIGNA, S.. **The webgraph framework i: compression techniques**. In: WWW '04: PROCEEDINGS OF THE 13TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, p. 595–602. ACM Press, 2004. 4, 4, 4.4

[Bol04b] BOLDI, P.; VIGNA, S.. **The webgraph framework ii: Codes for the world-wide web**. In: DCC '04: PROCEEDINGS OF THE CONFERENCE ON DATA COMPRESSION. IEEE Computer Society, 2004. 4, 4

[Bol09] BOLDI, P.; SANTINI, M. ; VIGNA, S.. **Permuting web graphs**. In: WAW '09: PROCEEDINGS OF THE 6TH INTERNATIONAL WORKSHOP ON ALGORITHMS AND MODELS FOR THE WEB-GRAPH, p. 116–126. Springer-Verlag, 2009. 4

[Bon05] BONATO, A.. **A Survey of Models of the Web Graph**. 2005. 4

[Bue08] BUEHRER, G.; CHELLAPILLA, K.. **A scalable pattern mining approach to web graph compression with communities**. In: WSDM '08: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON WEB SEARCH AND WEB DATA MINING, p. 95–106. ACM, 2008. 4, 4, 4.4, 4.4.1, 5

[Car90] CARLSSON, S.; LEVCOPOULOS, C. ; PETERSSON, O.. **Sublinear merging and natural mergesort**. In: PROCEEDINGS OF THE INTL. SYMPOSIUM ON ALGORITHMS, p. 251–260, 1990. 3.2.2

[Cas07] CASTILLO, C.; DONATO, D.; GIONIS, A.; MURDOCK, V. ; SILVESTRI, F.. **Know your neighbors: web spam detection using the web topology**. In: SIGIR '07: PROCEEDINGS OF THE 30TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, p. 423–430. ACM, 2007. 4, 4.3.2

[Cha99] CHAKRABARTI, S.; VAN DEN BERG, M. ; DOM, B.. **Focused crawling: a new approach to topic-specific web resource discovery**. In: WWW '99: PROCEEDING OF THE EIGHTH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, p. 1623–1640. Elsevier North-Holland, Inc., 1999. 4

[Chi2009] CHIERICHETTI, F.; KUMAR, R.; LATTANZI, S.; MITZENMACHER, M.; PANCONESI, A. ; RAGHAVAN, P.. **On compressing social networks**. In: PROCEEDINGS OF THE 15TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, p. 219–228. ACM New York, NY, USA, 2009. 1.1.1

[Chi09] CHIERICHETTI, F.; KUMAR, R.; LATTANZI, S.; PANCONESI, A. ; RAGHAVAN, P.. **Models for the Compressible Web**. lightless.org, 2009. 4, 5

[Cho98] CHO, J.; GARCIA-MOLINA, H. ; PAGE, L.. **Efficient crawling through url ordering**. In: WWW7: PROCEEDINGS OF THE SEVENTH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB 7, p. 161–172. Elsevier Science Publishers B. V., 1998. 4

[Chr78a] CHRISTEN, C.. **Improving the bound on optimal merging**. In: PROCEEDINGS OF THE 19TH IEEE SYMPOSIUM ON FOUNDATION OF COMPUTER SCIENCE), p. 259–266, 1978. 3.1.1

[Cla07] CLAUDE, F.; NAVARRO, G.. **A fast and compact web graph representation**. String Processing and Information Retrieval, p. 118–129, 2007. 4, 4.4

[Cov2006] COVER, T. M.; THOMAS, J. A.. **Elements of information theory**. John Wiley and Sons, 2006. 2

[Dea99] DEAN, J.; HENZINGER, M. R.. **Finding related pages in the world wide web**. In: WWW '99: PROCEEDINGS OF THE EIGHTH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, p. 1467–1479. Elsevier North-Holland, Inc., 1999. 4

[Dem00] DEMAINE, E. D.; LÓPEZ-ORTIZ, A. ; MUNRO, J. I.. **Adaptive set intersections, unions and differences**. In: PROCEEDINGS OF THE 11TH ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS (SODA), p. 743–752, 2000. 3.2.2

[Dou07] DOURISBOURE, Y.; GERACI, F. ; PELLEGRINI, M.. **Extraction and classification of dense communities in the web**. In: WWW '07: PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, p. 461–470. ACM Press, 2007. 4

[Dud81] DUDZINSKI, K.; DYDEK, A.. **On a stable storage merging algorithm**. Information Processing Letters, 12:5–8, 1981. 3, 3.1.1, 3.2.2, 5

[Eli75] ELIAS, P.. **Universal codeword sets and representations of the integers**. IEEE Transactions on Information Theory, IT-21(2):194–203, January 1975. 2.2.3, 2.2.5

[Epp1994] EPPSTEIN, D.. **Arboricity and bipartite subgraph listing algorithms**. Information Processing Letters, 51(4):207–211, 1994. 4.3.2

[Fla02] FLAKE, G. W.; LAWRENCE, S.; GILES, C. L. ; COETZEE, F. M.. **Self-organization and identification of web communities**. Computer, 35(3):66–70, 2002. 4

[Frw10] FRAMEWORK, W.. **version 2.4.4, http://webgraph.dsi.unimi.it/**, 2010. 4.4, 5

[Gal78] GALLAGER, R. G.. **Variations on a theme by Huffman**. IEEE Transactions on Information Theory, IT-24:668–674, November 1978. 2.2.3

[Gol66] GOLOMB, S. W.. **Run-length codings**. IEEE Transactions on Information Theory, 12(7):399–401, May 1966. 2.2.4, 2.2.5, 4.1.3

[Goo08] GOOGLE. **http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html**, 2008. 1.1.1, 4

[Gui02] GUILLAUME, J.-L.; LATAPY, M. ; VIENNOT, L.. **Efficient and simple encodings for the web graph**. p. 40–55. 2002. 4

[Hel94] HELLERSTEIN, J.; PFEFFER, A.. **The rd-tree: An index structure for sets**. Univ. of Wisconsin CS Technical Report, 1252, 1994. 4.2.2

[Hil2011] HILBERT, M.; LOPEZ, P.. **The world's technological capacity to store, communicate, and compute information**. Science, 2011. 1

[Huf52] HUFFMAN, D. A.. **A method for the construction of minimum-redundancy codes**. Proceedings of the IRE, p. 1098–1101, September 1952. 2.2.5

[Hwa72] HWANG, F. K.; LIN, S.. **A simple algorithm for merging two disjoint linearly ordered lists**. SIAM Journal of Computing, 1:31–39, 1972. 3, 3.1.1, 3.2.1, 5

[Kle99a] KLEINBERG, J. M.. **Authoritative sources in a hyperlinked environment**. Journal of the ACM, 46(5):604–632, 1999. 4, 4.3.1, 4.3.2

[Knu73] KNUTH, D. E.. **The Art of Computer Programming: Sorting and Searching**, volumen 3. Addison-Wesley, 1973. 3, 3.1.1, 3.2.1

[Kum99] KUMAR, R.; RAGHAVAN, P.; RAJAGOPALAN, S. ; TOMKINS, A.. **Trawling the web for emerging cyber-communities**. Comput. Networks, 31(11-16):1481–1493, 1999. 4, 4.3.2

[Mah06] MAHDIAN, A.; KHALILI, H.; NOURBAKHSH, E. ; GHODSI, M.. **Web graph compression by edge elimination**. In: DCC '06: PROCEEDINGS

OF THE DATA COMPRESSION CONFERENCE, p. 459+. IEEE Computer Society, 2006. 4

[Man79] MANACHER, G. K.. **Significant improvements to the Hwang-Lin merging algorithm**. Journal of ACM, 26:434–440, 1979. 3.1.1

[Mof97] MOFFAT, A.; BELL, T. C. ; WITTEN, I. H.. **Lossless compression for text and images**. International Journal of High Speed Electronics and Systems, 8(1):179–231, October 1997. 2.2.5

[Mof00] MOFFAT, A.; STUIVER, L.. **Binary interpolative coding for effective index compression**. Information Retrieval, 3(1):25–47, 2000. 3, 3.2.2, 5

[Mub2010] MUBAYI, D.; TURÁN, G.. **Finding bipartite subgraphs efficiently**. Information Processing Letters, 110(5):174–177, 2010. 4.3.2

[Naj05] NAJORK, M.; WIENER, J. L.. **Breadth-first crawling yields high-quality pages**. In: WWW '01: PROCEEDINGS OF THE 10TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, p. 114–118. ACM Press, 2001. 4

[Nto04] NTOULAS, A.; CHO, J. ; OLSTON, C.. **What's new on the web?: the evolution of the web from a search engine perspective**. In: WWW '04: PROCEEDINGS OF THE 13TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, p. 1–12. ACM Press, 2004. 4

[Pag98] PAGE, L.; BRIN, S.; MOTWANI, R. ; WINOGRAD, T.. **The pagerank citation ranking: Bringing order to the web**. Technical report, Stanford Digital Library Technologies Project, 1998. 4, 4.3.1

[Rag03] RAGHAVAN, S.; GARCIA-MOLINA, H.. **Representing web graphs**. In: DATA ENGINEERING, 2003. PROCEEDINGS. 19TH INTERNATIONAL CONFERENCE ON, p. 405–416, 2003. 4, 4, 4.3.2

[Ran02] RANDALL, K. H.; STATA, R.; WICKREMESINGHE, R. G. ; WIENER, J. L.. **The link database: fast access to graphs of the web**. In: DATA COMPRESSION CONFERENCE, 2002. PROCEEDINGS. DCC 2002, p. 122–131, 2002. 4

[Ric79] RICE, R. F.. **Some practical universal noiseless coding techniques**. Technical Report 79-22, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, 1979. 2.2.4, 2.2.5, 3, 3.2.3, 5

[Ris79] RISSANEN, J. J.; JR., G. G. L.. **Arithmetic coding**. IBM Journal of Research and Development, 23(2):146–162, March 1979. 2.2.5

[Sha48] SHANNON, C. E.. **A mathematical theory of communication**. The Bell System Technical Journal, 27:379–423, 623–656, July 1948. 1

[Sto80] STOCKMEYER, P. K.; YAO, F. F.. **On the optimality of linear merge**. SIAM Journal of Computing, 9:85–90, 1980. 3.1.1, 3.1.2, 3.2.3

[Sue01] SUEL, T.; YUAN, J.. **Compressing the graph structure of the web**. In: DCC '01: PROCEEDINGS OF THE DATA COMPRESSION CONFERENCE (DCC '01). IEEE Computer Society, 2001. 4, 4, 4.4, 5

[Veg93] DE LA VEGA, W. F.; KANNAN, S. ; SANTHA, M.. **Two probabilistic results on merging**. SIAM Journal of Computing, 22(2):261–271, April 1993. 3, 3.1.1, 3.2.3, 3.2.3, 3.2.3, 5

[Wat98] WATTS, D. J.; STROGATZ, S. H.. **Collective dynamics of 'small-world' networks.** Nature, 393(6684):440–442, June 1998. 2

[Web10] DATASET, W.. **http://law.dsi.unimi.it/**, 2010. 4.4

A
Desigualdade da Entropia

Teorema A.1 *Se a e b são inteiros tal que $a, b \geq 1$ e $H(a/(a+b), b/(a+b))$ é a função de entropia binária então:*

$$(a+b)H\left(\frac{a}{a+b}, \frac{b}{a+b}\right) - \log_2 \binom{a+b}{a} \geq 1.$$

Prova. Simplificando a parte esquerda da desigualdade, temos que:

$$\begin{aligned} &= (a+b)H\left(\frac{a}{a+b}, \frac{b}{a+b}\right) - \log_2 \binom{a+b}{a} \\ &= \log_2 \frac{(a+b)^{a+b}}{a^a b^b} \cdot \frac{a!b!}{(a+b)!} \\ &= \log_2 B(a,b). \end{aligned}$$

Podemos focar em mostrar, por indução, que $B(a,b) \geq 2$. Note que $B(a,b) = B(b,a)$. A idéia desta prova é mostrar que $B(a,b+1) \geq B(a,b) \geq 2$ fixando a e assumindo que $a \geq b$. Começamos mostrando o caso base $B(a,a) \geq 2$, também por indução.

Para $a = 1$, temos que:

$$B(1,1) = \frac{(1+1)^{1+1}}{1^1 1^1} \cdot \frac{1!1!}{(1+1)!} = 2 \geq 2.$$

Simplificando $B(a,a)$, temos que:

$$B(a,a) = \frac{(a+a)^{a+a}}{a^a a^a} \cdot \frac{a!a!}{(a+a)!} = 2^{2a} \cdot \frac{(a!)^2}{(2a)!}.$$

Seguimos a mesma estratégia e mostramos que $B(a, a+1) \geq B(a, a) \geq 2$:

$$\begin{aligned} B(a, a+1) &\geq B(a, a) \\ 2^{2a+2} \cdot \frac{((a+1)a!)^2}{(2a+2)!} &\geq 2^{2a} \cdot \frac{(a!)^2}{(2a)!} \\ 2^2 \cdot \frac{(a+1)^2}{(2a+2)(2a+1)} &\geq 1 \\ \frac{2a+2}{2a+1} &\geq 1. \end{aligned}$$

O caso base está completo e seguimos para o passo da indução. Assumimos que $a \geq b$, fixe a e, por indução em b, mostramos que $B(a, b+1) \geq B(a, b) \geq 2$:

$$\begin{aligned} B(a, b+1) &\geq B(a, b) \\ \frac{(a+b+1)^{a+b}}{a^a(b+1)^b} \cdot \frac{a!b!}{(a+b)!} &\geq \frac{(a+b)^{a+b}}{a^a b^b} \cdot \frac{a!b!}{(a+b)!} \\ \frac{(a+b+1)^{a+b}}{(b+1)^b} &\geq \frac{(a+b)^{a+b}}{b^b} \\ \left(\frac{a+b+1}{a+b}\right)^{a+b} &\geq \left(\frac{b+1}{b}\right)^b \\ \left(1+\frac{1}{a+b}\right)^a &\geq \left(\frac{b+1}{b} \cdot \frac{a+b}{a+b+1}\right)^b \\ \left(1+\frac{1}{a+b}\right)^a &\geq \left(1+\frac{1}{b(a+b+1)}\right)^b. \end{aligned}$$

Como $a+b \leq b(a+b+1)$ e $a \geq b$, a demonstração está completa. ■