به نام آن که هیچ رمزی برایش پوشیده نیست

پاسخ سؤال ۶ تمرین سری ۱

حسین هادی‌پور
دانشکده ریاضی آمار و علوم کامپیوتر دانشگاه تهران

از آن‌جایی که احساس کردم برخی از دوستان، سؤال ۶ که در رابطه با ترتیب فشرده سازی و رمزنگاری بود را به خوبی متوجه نشدند ،ترجیح دادم پاسخ آن را به صورت مجزا بیاورم.

پاسخ ۶ اول فشرده می‌کنم سپس رمز می‌کنم!

دلیل این کار در تئوری اطلاعات که توسط شانون پایه‌گذاری شد نهفته است. در مخابرات دو نوع کدینگ وجود دارد:

۱. کدینگ منبع
۲. کدینگ کانال

بحث فشرده سازی به کدینگ منبع برمی‌گردد. فرض کنید داده‌ای که قرار است فشرده-رمز کنیم خروجی یک منبع گسسته بدون حافظه M نمادی باشد. در کدینگ منبع، دو هدف را پیش روی خود قرار می دهیم:

۱. تعداد بیت‌های مورد نیاز بر واحد زمان که برای نمایش خروجی منبع لازم است حداقل باشد.
۲. بازسازی(دیکدینگ) توسط دنباله‌ی باینری بدون ابهام امکان پذیر باشد.

در واقع فشرده سازی هدف اول ما در کدینگ منبع است. طول متوسط کد را به صورت زیر تعریف می‌کنیم: $\overline{N} = \sum_{i=۱}^{M} N_i P_i$ که در آن N_i و P_i به ترتیب طول و احتمال رخداد سمبل i ام هستند.
طبق قضیه کدینگ منبع شانون طول متوسط کدگذاری دارای کران‌های زیر است:

$$H(x) \leq \overline{N} \leq H(x) + \epsilon$$

که در آن ϵ یک مقدار مثبت است و $H(x)$ آنتروپی منبع یا داده است و از رابطه زیر بدست می‌آید:

$$H(x) = \sum_{i=۱}^{M} P_i log_۲ \frac{۱}{P_i}$$

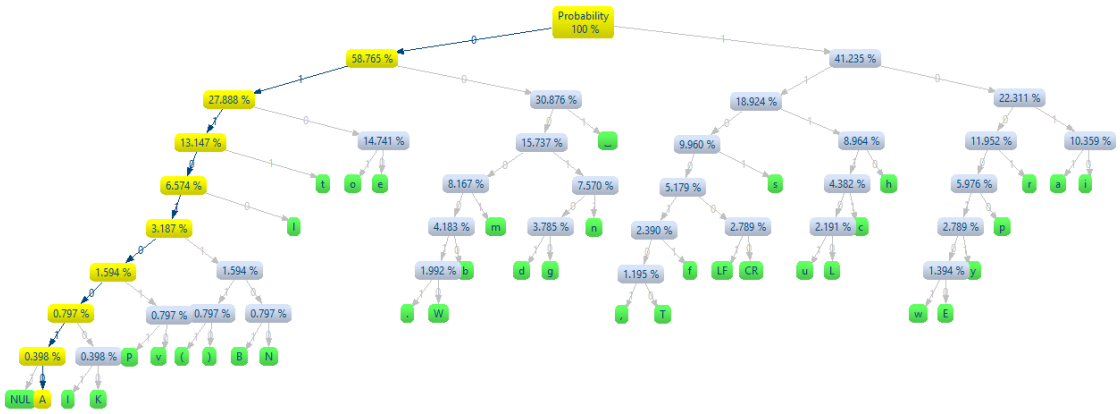
در واقع $H(x)$ امید ریاضی متغیر تصادفی $I_i = \frac{۱}{P_i}$ یا همان اطلاعات سمبل i ام است. اگر کدینگ منبع بهینه باشد ثابت می‌شود که باید طول کدها در رابطه زیر صدق کند:

$$N_i = [log_۲ \frac{۱}{P_i}]$$

در نتیجه سمبل‌هایی که احتمال رخداد آن‌ها زیاد باشد طول کمتر و سمبل‌هایی که احتمال وقوع آن‌ها کم باشد طول بیشتری خواهند داشت و این راز فشرده سازی است. برای مثال در کدینگ هافمن که تقریبا بهینه است و برای فشرده سازی استفاده می‌شود از همین ایده استفاده می‌شود و سمبل‌های پر فرکانس نسبت به سمبل‌های نادر طول کمتری نصیبشان می‌شود. برای این که مطلب روشن‌تر شود یک متن لاتین را در زیر با الگوریتم هافمن کد کرده‌ایم:

The Learning With Error problem (LWE) is becoming more and more used in cryptography, for instance, in the design of some fully homomorphic encryption schemes. It is thus of primordial importance to find the best algorithms that might solve this problem so that concrete parameters can be proposed. The BKW algorithm was proposed by Blum et al. as an algorithm to solve the Learning Parity with Noise problem (LPN), a subproblem of LWE. This algorithm was then adapted to LWE by Albrecht et al.

در زیر درخت هافمن که در الگوریتم هافمن بدست می‌آید نشان داده شده:



شکل ۱: درخت هافمن

همان طور که مشاهده می‌شود سمبل‌هایی که احتمال وقوع آن‌ها کمت است کد باینری نسبت داده شده به آن‌ها طویل‌تر است و بالعکس. طول متوسط کد در این حالت برابر است با ۴/۵۵.

اما در رمزنگاری احتمال رخداد سمبل‌ها به سمت یکنواخت شدن میل می‌کند! به عبارت دیگر خاصیت الگوریتم‌های رمزنگاری این است که توزیع فرکانسی ورودی را به صورت یکنواخت در خروجی پخش می‌کنند و به همین خاطر است که ما ابتدا فشرده سازی را انجام می‌دهیم و سپس رمز می‌کنیم زیرا در غیر این صورت فشرده سازی که مبنی بر تفاوت فرکانس نسبی سمبل‌ها است، به خوبی عمل

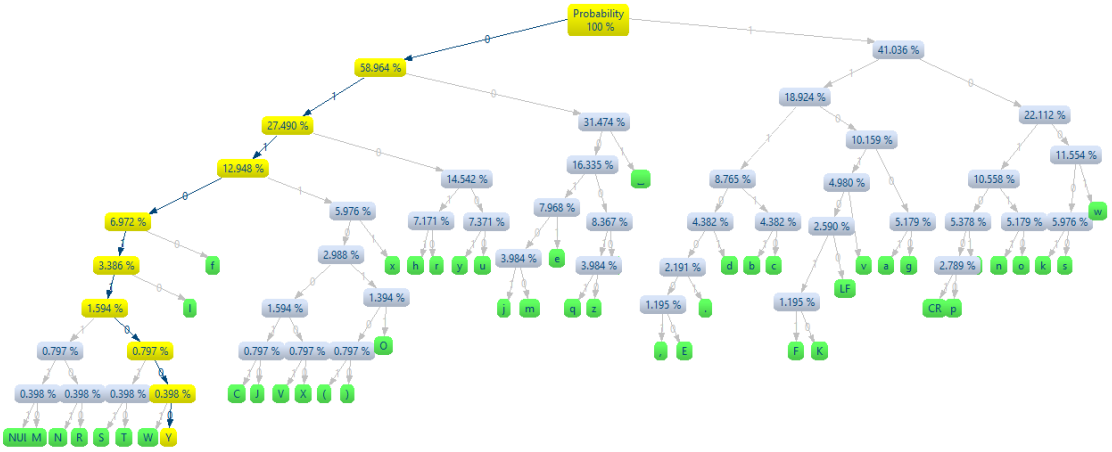
نمی‌کند. به زبان نظریه اطلاعات، پس از رمزنگاری آنتروپی آفزایش می‌یابد زیرا آنتروپی یک منبع گسسته بدون حافظه M نمادی دارای کران زیر است.

$$H(X) \leq log_۲ M$$

و تساوی زمانی رخ خواهد داد که توزیع احتمال سمبل‌ها یکنواخت باشد. لذا طبق رابطه‌ای که بین طول متوسط کد و آنتروپی وجود دارد طول متوسط هم افزایش خواهد یافت و ضریب فشرده سازی نیز کاهش می‌یابد. در زیر متن لاتینی که در ابتدا آوردیم را با الگوریتم هیل رمز کرده‌ایم:

Vde Fakhzkpk Ogki Wiyue ehubtiw (FOE) gp ealzabic vrno uhn ksot gtbb gt iitrfsuktnha, bux ycfjubyt, dr nzq enctlp wh gvfm ozxfu neeefnhpfyf njirawuwhy enysbvc. Ty ob wrgy ai czmygwssty ijeqyimqty or vcfv hre wrwu jdaufifvao zjav eygag udevk lxkw fxwvdsu hp zjah uyjiriqh rwhejxdexy rrd rg fxwjsgoo. Cda NMJ hdaufifru sqx wtmpyfrn hs Ffiw pw cy. vm ji wcboixnrm xk eekws skm Kxmwwkut Rwhcez qlwf Xiqrh fxwvdsq (KEJ), s wkguhubtiw eo YSV. Exkw wcboixnrm inl vdel ihusgoo cq CXK sl Oqwlywzx ml kz.

لازم به ذکر است که آنتروپی متن فوق قبل از رمزنگاری برابر با ۴/۱۴و پس از رمزنگاری برابر با ۴/۶۲ است. درخت هافمن را برای متن فوق در زیر مشاهده می‌کنیم:



شکل ۲: درخت هافمن متن رمزشده

طول متوسط کد در این حالت برابر با ۴/۹۸ است!

با این که سیستم رمزنگاری هیل یک سیستم کلاسیک است وتوزیع احتمال سمبل‌ها را به خوبی به سمت یکنواخت نمی‌برد، مشاهده می‌شود که آنتروپی افزایش یافت و طول متوسط هم افزایش یافت ،که به دنبال آن فشرده سازی به خوبی انجام نشد.