بررسی مدل ترای‌های d-d یی اریب تصادفی

رامین کاظمی*، حدیثه عبدالهی نهوجی؛ دانشگاه بین‌المللی امام خمینی(ره)، گروه آمار
سولماز نوروزی؛ دانشگاه غیرانتفاعی البرز، گروه آمار

دریافت ۹۴/۱۰/۲۱ پذیرش ۹۵/۴/۲

چکیده

ترای‌ها عمومی‌ترین ساختار داده‌ای روی رشته‌ها هستند. با استفاده از رشته‌ها روی الفبایی که منجر به تولید درخت‌های d-dیی می‌شود، می‌توان ترای‌های d-dیی ساخت. در سراسر مقاله فرض می‌کنیم که رشته‌های ذخیره شده در ترای به‌کمک منشأ بی‌حافظه مناسب تولید می‌شوند. در این مقاله، تحلیل میانگین نمایه با رهیافت ترکیبیاتی خاصی به ترای‌های d-d یی توسیع داده می‌شود. از این رهیافت ترکیبیاتی برای بررسی میانگین نمایه استفاده می‌کنیم زیرا تابع احتمال آن نامعلوم است. تابع احتمال عمق و تابع توزیع ارتفاع را هنگامی که n بزرگ است، به‌دست می‌آوریم. این نتایج از بررسی معادله‌های بازگشتی مشخصی که آن‌ها را با روش تحلیلی حل می‌کنیم، به‌دست می‌آیند.

واژه‌های کلیدی: ترای‌های d-d یی، نمایه، ارتفاع، عمق.

مقدمه

درخت‌های رقومی، ساختارهای داده‌ای بنیادی روی رشته‌ها هستند [۱۳]. در میان آن‌ها ترای و درخت جستجوی رقومی به‌سبب کاربردهای وسیع آن‌ها بیش‌تر از سایرین به آن توجه شده است. ترای‌ها از اولین ساختارهای داده‌ای مفید برای اهداف بازیابی، فهرست‌سازی و شاخص‌گذاری محسوب می‌شوند. این ساختارهای درختی را اولین بار بریاندایس در اواخر دهۀ ۵۰ میلادی برای پردازش اطلاعات معرفی کرد [۱]. در سال ۱۹۶۰، فردکین نام اخیر را برای این نوع از درخت‌های رقومی پیشنهاد کرد [۷]. ترای‌ها درخت‌های چندشاخه‌ای هستند که گره‌هایشان شامل گرافیک بردارهایی (شکل حروف در یک نوشتار رسم شده به‌کمک خم‌ها و خطوط) از نوشته‌ها یا رقم‌هاست. به‌دلیل سادگی و کارایی، ترای‌ها به‌سرعت در کاربردهای گوناگون استفاده شدند. امروزه آن‌ها از رده‌بندی اسناد تا مراجعه به آدرس‌های IP، از فشرده‌سازی داده‌ها تا خرد کردن حافظه با دست‌یابی تصادفی و با محتوایی که مرتب به‌هنگام می‌شود، از الگوریتم انتخاب اولیۀ اطلاعات تا جدول‌های خرد کنندۀ بخش شده به‌کار گرفته می‌شوند [۸]. ترای‌ها به‌وفور در بسیاری از کاربردهای علوم رایانه‌ای استفاده می‌شوند. برای مثال ترای‌ها به‌طور گسترده‌ای در الگوریتم‌ها برای تصحیح خودکار کلمات در متن‌ها [۱۲] و در الگوریتم‌ها برای رده‌بندی و رفع اشکال برنامه‌ها به‌کمک توابع استفاده می‌شوند. هم‌چنین ترای‌ها در جستجو، مرتب‌کردن، کدگذاری، طرح فشرده‌سازی لمپل-زیو و زیست‌شناسی مولکولی کاربرد دارند [۱۲]، [۱۵]. در حقیقت ترای‌ها انتخابی طبیعی از ساختارهای داده‌ای هستند که ورودی‌های آن‌ها شامل عباراتی الفبایی یا رقمی است. آن‌ها اغلب برای ذخیره کردن چنین داده‌هایی به‌منظور بازیابی کارا استفاده می‌شوند.

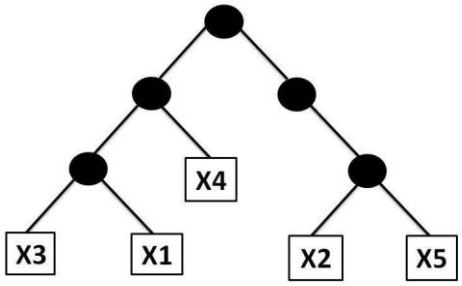
*نویسنده مسئول r.kazemi@sci.ikiu.ac.ir

فرض کنید n رشتهٔ ساخته شده روی الفبای متناهی $\{a_1,...,a_d\}$ مفروض است، که در آن $d \geq 2$. یک ترای از رشته‌های بالا به‌صورت زیر ساخته می‌شود. اگر $n=0$، آن‌گاه ترای تهی است. اگر $n=1$، آن‌گاه ترای یک گره (داخلی) یکتا برای در بر گرفتن این رشته است. اگر $n \geq 2$، آن‌گاه یک گره داخلی به‌عنوان ریشه در نظر گرفته می‌شود و سایر رشته‌ها براساس اولین رقم‌شان به یکی از d زیردرخت متصل به این گره برای ذخیره‌شدن از چپ به‌راست هدایت می‌شوند. این الگوریتم برای ذخیره شدن رشته‌ها در زیردرخت‌ها به‌طور بازگشتی به‌کار گرفته می‌شود. به این‌صورت که $\ell+1$ امین رقم هر رشته برای شاخه گزینی در سطح ℓ اهمیت دارد. در نهایت هر رشته در یک برگ ذخیره می‌شود [۳]. برای مثال فرض کنید $d=2$ و رشته‌های زیر مفروض باشند:

$$X_1 = 00111...,$$
$$X_2 = 11011...,$$
$$X_3 = 00011...,$$
$$X_4 = 01010...,$$
$$X_5 = 11111...,$$

این الگوریتم بیان می‌کند که رشته‌های X_4, X_3, X_1 به زیردرخت سمت چپ و X_2, X_5 به زیردرخت سمت راست برای ذخیره شدن هدایت می‌شوند زیرا اولین رقم رشته های X_4, X_3, X_1؛ صفر و اولین رقم رشته‌های X_2, X_5، یک است. هم‌چنین در زیردرخت چپ رشته‌های X_1, X_3 به‌سمت چپ و رشتهٔ X_4 به‌سمت راست هدایت می‌شود. در مقابل و در زیردرخت راست هر دو رشتهٔ X_2, X_5 به‌سمت راست هدایت می‌شوند. زیردرخت شامل X_1, X_3 خود به یک زیردرخت سمت چپ با X_3 و یک زیردرخت سمت راست با X_1 تفکیک می‌شود. چون X_3 تنها رشته برای ذخیره شدن در زیردرخت سمت چپ و X_1 نیز تنها رشته برای ذخیره شدن در زیردرخت سمت راست است، از این رو، این زیردرخت تنها شامل دو گره خارجی (برگ) برای ذخیره شدن این دو رشته است. به‌همین صورت سه رشتهٔ دیگر در درخت ذخیره می‌شوند. ترای حاصل در شکل ۱ رسم شده است که در آن دایره‌ها بیان‌گر گره‌های داخلی و مربع‌ها بیان‌گر گره‌های خارجی هستند [۴]. مهم‌ترین متغیر تصادفی در بررسی درخت‌های تصادفی و به‌ویژه ترای، نمایه است زیرا متغیرهای دیگری چون ارتفاع، عمق یک گره و کوتاه‌ترین مسیر را می‌توان بر حسب آن نوشت [۹]. این متغیر شامل دو نوع افقی و عمودی است. نمایه افقی، تابعی از تعداد رشته‌های ذخیره شده و فاصله از ریشه است [۱۴]. چون گره‌هایی که رشته‌ها در آن‌ها ذخیره می‌شوند (گره‌های خارجی) از اهمیت بیش‌تری در مقایسه با گره‌های داخلی برخوردارند. از این رو، در این مقاله نمایهٔ افقی خارجی بحث می‌شود. نتایج به سادگی و با اندکی تغییرات به نمایه افقی داخلی تعمیم داده می‌شوند. از نماد $B_{n,k}$ برای نمایه افقی خارجی یعنی تعداد گره‌های خارجی در فاصله k (سطح k) هنگامی که n رشته در ترای ذخیره می‌شود، استفاده می‌کنیم. موقعیت عمودی یک گره برابر تفاضل تعداد گام‌های به‌راست و چپ روی مسیر از ریشه تا آن گره است. نمایهٔ عمودی برابر تعداد گره‌های موجود در یک موقعیت عمودی است و مانند نمایه افقی برحسب گره‌های داخلی و خارجی بر دو نوع است. در این مقاله، نمایه افقی خارجی ساخته شده روی n رشتهٔ d نوعی تولید شده با منشأ بی‌حافظه بررسی می‌شود. در این مدل هر رشته یک دنبالهٔ d نوعی مستقل و هم‌توزیع با احتمال $0 < p_d < p_{d-1} < .... < p_1 < 1$ برای هر کدام از نوع‌ها است $(\sum_{i=1}^{d} p_i = 1)$. واضح است که برای هر $d \geq 2$، مجموعی از p_i ها کم‌تر از مجموع دیگر آن‌ها است. در هر نابرابری، تعداد p_i های در سمت

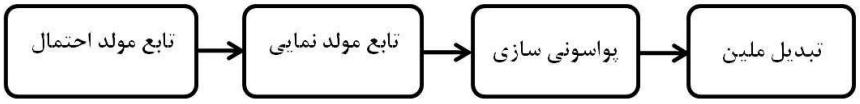
چپ را با t، حاصل جمع سمت چپ را با $p_1(t)$ و حاصل جمع سمت راست را با $p_2(t)$ نشان می‌دهیم. تحلیل در این مقاله برای هر t که نابرابری را برقرار می‌کند، اعتبار دارد. این موضوع تفاوت اصلی رهیافت ترکیبیاتی استفاده شده در این مقاله با سایر رهیافت‌های ترکیبیاتی است.



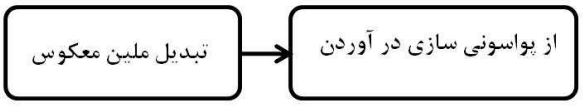
شکل ۱. ترای ساخته شده از ۵ رشتهٔ $X_1,...,X_5$.

فرض کنید $1 \le s \le d$ ای وجود داشته باشد به‌طوری‌که $p_d + p_{d-1} + ... + p_s < p_{s-1} + p_{s-2} + ... + p_1$.

برای مثال اگر $0 < p_3 = 0/1 < p_2 = 0/3 < p_1 = 0/6 < 1$، آن‌گاه $s = 1,2$ و تحلیل برای $p_3 < p_2 + p_1$ و $p_3 + p_2 < p_1$ برقرار است. این ویژگی که تعمیمی از منشأ بی‌حافظه برای رشته‌های دودویی است یک ترای اریب را ایجاد می‌کند.



شکل ۲. گام های ترتیبی یک تحلیل ترکیبیاتی استاندارد



شکل ۳. بازیافت گشتاورها

چون محاسبهٔ توزیع احتمال متغیرهای تصادفی تعریف شده روی الگوریتم‌ها یا درخت‌های تصادفی اغلب غیرممکن است تحلیل ترکیبیاتی می‌تواند با اجرای چند گام ترتیبی بدون ارجاع به توزیع احتمال متغیر تصادفی منجر به پیدا کردن میانگین یا واریانس این متغیرهای تصادفی شود. در برخی موارد از جمله بحث حاضر تنها یک تحلیل مجانبی امکان‌پذیر است. به‌طورکلی یک تحلیل ترکیبیاتی استاندارد شامل مراحل نشان داده شده در شکل ۲ است. اگر هدف انجام تحلیلی مجانبی باشد، آن‌گاه چون انتگرال در تبدیل ملین به یک پارامتر بزرگ (در این‌جا تعداد رشته‌ها) وابسته خواهد شد یک تحلیل نقطهٔ زینی نیز ضروری است [۱۵]. برای بازیافت ضرایب تبدیل پواسن در این روش که در این‌جا گشتاورهای مرتبهٔ اول نمایهٔ افقی خارجی هستند باید گام‌های ترتیبی نمایش داده شده در شکل ۳ انجام شود [۱۴].

در بخش ۲ ابتدا نشان داده می‌شود که تابع مولد احتمال نمایه افقی خارجی در رابطه‌ای بازگشتی صدق می‌کند. شرایط آغازین این معادله با توجه به تعریف نمایه به‌دست می‌آیند. چون ترای یک ساختار برچسب‌دار است با تعریف یک تابع مولد نمایی مناسب و استفاده از آن، رابطه‌ای بازگشتی برای تبدیل پواسن متناظر به دست می‌آید. در بخش ۳

تبدیل مِلین که بر اساس تبدیل پواسن تعریف می‌شود بررسی شده و ثابت می‌شود که در رابطه‌ای بازگشتی صدق می‌کند. در بخش ۴ دامنه‌های مختلفی برای تحلیل ترکیبیاتی از طریق تحلیل نقطهٔ زینی یک تابع خاص در نظر گرفته می‌شود. در ادامه نتایج اصلی را در قضیهٔ ۱ بیان و ثابت می‌کنیم. این نتایج با برگشت از پواسن سازی حاصل می‌شوند. با توجه به ارتباط میان نمایه افقی خارجی با عمق گره و ارتفاع ترای در بخش آخر به‌ترتیب تابع احتمال عمق گره و تابع توزیع ارتفاع ترای ارائه می‌شوند.

پواسن‌سازی

با توجه به تعریف نمایه افقی خارجی

$$B_{n,k}=\begin{cases}0, & n=0,k\geq 0\\ 1, & n=1,k=0\\ 0, & n=1,k\geq 1\\ 0, & n\geq 1,k=0\end{cases}$$

تابع مولد احتمال نمایه افقی خارجی یعنی $\phi_{n,k}(u)=E(u^{B_{n.k}})$ در رابطهٔ بازگشتی

$$j_{n,k}(u)=\sum_{k_i\geq 0}^{n}\binom{n}{k_1,\ldots,k_{d-1}}p_1^{k_1}\ldots p_{d-1}^{k_{d-1}}p_d^{n-\sum_{i=1}^{d-1}k_i}\times j_{k_1,k-1}(u)\ldots j_{k_{d-1}}(u)j_{n-\sum_{i=1}^{d-1}k_i,k-1}(u) \qquad (۱)$$

صدق می‌کند، که در آن $n\geq d$ و $k\geq ۱$. شرایط آغازین این معادله بدین‌صورت هستند:

$$\phi_{n,k}(u)=\begin{cases}1, & n=0,k\geq 0\\ u, & n=1,k=0\\ 1, & n=1,k\geq 1\\ 1, & n\geq 1,k=0\end{cases}$$

برای اثبات رابطهٔ (۱) کافی است به این نکته توجه کنیم که یک ترای دارای یک ساختار بازگشتی است. در واقع هر کدام از زیردرخت‌های ریشه یک ترای با ارتفاع یکی کم‌تر از ترای اصلی هستند و به‌دلیل فرض استقلال در منشأ بی‌حافظه تعداد رشته‌هایی که رقم اول آن‌ها یکی از انواع رقم‌هاست دارای توزیع چندجمله‌ای با پارامترهای $n, p_1,\ldots, p_{d-1}$است. چون ترای یک ساختار برچسب‌دار است تابع مولد نمایی (۲) را که در آن دنبالهٔ شمارشی، تابع مولد احتمال $\phi_{n,k}(u)$ است، در نظر می‌گیریم:

$$\mathrm{E}_k(x,u)=\sum_{n\geq 0}\phi_{n.k}(u)\frac{x^n}{n!} \qquad (۲)$$

در این‌صورت به‌کمک رابطهٔ (۱) می‌توان نتیجه گرفت:

$$\mathrm{E}_k(x,u)=\prod_{i=1}^{d}\mathrm{E}_{k-1}(p_i x,u)+(\varphi_{1,k}(u)-\varphi_{1,k-1}(u))x\ , \qquad k\geq 1.$$

از طرفی $\mathrm{E}_{\circ}(x,u)=e^x+x(u-1)$ و

$$\mathrm{E}_1(x,u)=\prod_{i=1}^{d}\left(e^{p_i x}+p_i x(u-1)\right)+(1-u)x$$

$$= e^{x} + \sum_{i\neq j}^{d} e^{p_i x} p_j x(u-1) + \sum_{i\neq j}^{d} p_i p_j x^2 (u-1)^2 + (1-u)x.$$

در نتیجه

$$\mathrm{E}_k(x,u) = \prod_{i=1}^{d} \mathrm{E}_{k-1}(p_i x, u) \quad k \geq 2. \qquad (۳)$$

با مشتق گرفتن از معادلهٔ (۳) نسبت به u و سپس قراردادن $u=1$ داریم:

$$\begin{aligned}\mathrm{E}_k(x) = {} & \mathrm{E}_{k-1}(p_1 x)\exp\{p_2 x + p_3 x + ... + p_d x\} \\ & + \mathrm{E}_{k-1}(p_2 x)\exp\{p_1 x + p_3 x + ... + p_d x\} \\ & + ... + \mathrm{E}_{k-1}(p_{d-1} x)\exp\{p_1 x + ... + p_{d-2} x + p_d x\} \\ & + \mathrm{E}_{k-1}(p_d x)\exp\{p_1 x + p_2 x + ... + p_{d-1} x\} \qquad (۴)\end{aligned}$$

که در آن

$$\mathrm{E}_k(x) = \sum_{n\geq 0} \mathrm{E}(B_{n,k}) \frac{x^n}{n!}, \quad k \geq 2.$$

پواسن‌سازی فنی است که مسئله را به‌سمت یک فرایند پواسن هدایت می‌کند. تبدیل پواسن، دنبالهٔ توصیف کنندهٔ مدل برنولی را درون تابع مولد یک متغیر مختلط توصیف کنندهٔ مدل پواسن می‌نگارد. هر زمان مسئله در قلمرو پواسن حل شود، باید به‌منظور رسیدن به نتیجهٔ اصلی از پواسن‌سازی در آید. این فن به‌ویژه برای توابع مولدی که تام هستند (در این‌جا، $E_k(x)$) به‌کار گرفته می‌شود. حال از تبدیل پواسن $P^{(k)}(x) = e^{-x} E_k(x)$ به‌منظور هدایت کردن مسئله به‌سمت یک فرایند پواسن استفاده می‌کنیم. چون $e^{-x} = e^{-(p_1+p_2+....+p_d)x}$، معادلهٔ (۴) به معادلهٔ (۵):

$$P^{(k)}(x) = \sum_{i=1}^{d} P^{(k-1)}(p_i x), \quad k \geq 2 \qquad (۵)$$

با شرایط آغازین $P^{(0)}(x) = x e^{-x}$ و

$$\begin{aligned}P^{(1)}(x) &= e^{-x}\left[\frac{\partial \mathrm{E}_1(x,u)}{\partial u}\right]_{u=1} \\ &= \sum_{i\neq j}^{d} e^{-(1-p_i)x} p_j x - x e^{-x}.\end{aligned}$$

برمی‌گردد.

تبدیل ملین

فرض کنید $P^{(k)*}(s)$ تبدیل ملین $P^{(k)}(x)$ باشد، یعنی

$$P^{(k)*}(s) = \int_0^{\infty} x^{s-1} P^{(k)}(x) dx, \quad s \in C,$$

در این‌صورت اولین نتیجه، لم ۱ را بیان و ثابت می‌کنیم:

لم ۱. برای هر $k \geq 2$،

$$P^{(k)*}(s) = \Gamma(s+1) T_d(s)^{k-1} \left(\sum_{i\neq j}^{d} p_j (1-p_i)^{-s-1} - 1 \right),$$

که در آن $T_d(s)=p_1^{-s}+\ldots+p_d^{-s}$.

برهان. با توجه به تعریف تبدیل ملین، معادلهٔ (۵) به معادلهٔ (۶) برگردانده می‌شود:

$$P^{(k)^*}(s)=(p_1^{-s}+\ldots.+p_d^{-s})P^{(k-1)^*}(s),\quad k\geq 2 \qquad (۶)$$

از طرفی

$$P^{(k)^*}(s)=\Gamma(s+1)\left(\sum_{i\neq j}^{d}p_j(1-p_i)^{-s-1}-1\right).$$

بنابراین با تکرار معادلهٔ (۶)،

$$P^{(k)^*}(s)=\Gamma(s+1)T_d(s)^{k-1}\left(\sum_{i\neq j}^{d}p_j(1-p_i)^{-s-1}-1\right),\qquad k\geq 2$$

و برهان کامل می‌شود.

نتایج اصلی

برای عدد حقیقی α با شرط

$$\left(-\log p_1(t)\right)^{-1}<\alpha<\left(-\log p_2(t)\right)^{-1}$$

قرار می‌دهیم:

$$\rho=\rho(\alpha)=\frac{1}{\log\frac{p_1(t)}{p_2(t)}}\log\left(\frac{1-\alpha\log\frac{1}{p_1(t)}}{\alpha\log\frac{1}{p_2(t)}-1}\right).$$

به‌طور هم‌ارز

$$\alpha=\frac{T_d(\rho)}{p_1(t)^{-\rho}\log\frac{1}{p_1(t)}+p_2(t)^{-\rho}\log\frac{1}{p_2(t)}}.$$

به‌علاوه فرض می‌کنیم:

$$\beta(\rho)=\frac{p_1(t)^{-\rho}\,p_2(t)^{-\rho}\log\left(\frac{p_1(t)}{p_2(t)}\right)^2}{T_d(\rho)^2}.$$

برای چگونگی تمایز بین دامنه‌های مختلف در بررسی نمایه افقی خارجی و برخی جزئیات محاسبه‌ای در قسمت‌های بعدی، خوانندگان به بررسی دقیق مقالهٔ کاظمی و وحیدی اصل [۹]، [۱۱] ارجاع داده می‌شوند. به هر حال در این بررسی نیز مجبور به تمایز بین چند دامنه روی حوزهٔ تغییرات α هستیم. تحلیل را با دامنهٔ:

$$\left(-\log p_1(s)\right)^{-1}<\alpha=k/\log n<\frac{T_d(-2)}{-p_1(t)^2\log p_1(t)-p_2(t)^2\log p_2(t)}$$

شروع می‌کنیم. در این حالت $-2<\rho_{n,k}=\rho(k/\log n)<\infty$ و می‌توان ρ را $\rho_{n,k}$ انتخاب کرد (تحلیل نقطهٔ زینی). از این رو، بنا بر لم ۱ و تبدیل ملین وارون [۱۵] داریم:

$$P^{(k)}(x)=\frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty}\Gamma(s+1)\frac{\left(\sum_{i\neq j}^{d}p_j(1-p_i)^{-s-1}-1\right)}{T_d(s)}T_d(s)^k x^{-s}ds, \qquad (۷)$$

که در آن $\rho>-2$. برای تحلیل مجانبی انتگرال در (۷) طبیعی است که $\rho=\rho_{n,k}$ که رابطهٔ (۸) را برقرار می‌سازد:

$$\frac{k}{\log n}=\frac{T_d(\rho)}{p_1(t)^{-\rho}\log\frac{1}{p_1(t)}+p_2(t)^{-\rho}\log\frac{1}{p_2(t)}} \qquad (۸)$$

و به‌عنوان نقطهٔ زینی تابع $T_d(s)^k n^{-s}=e^{k\log T_d(s)-s\log n}$ انتخاب شود، یعنی $\rho=\rho_{n,k}=\rho(k/\log n)$. چون $|T_d(s_j)|=T_d(\rho)$، بنا بر این روی خط $\Re(s)=p$ تعداد نامتناهی نقطهٔ زینی وجود دارند [۹]:

$$s_j=\rho+\frac{2\pi ij}{\log\frac{p_1(t)}{p_2(t)}}$$

در نتیجه رفتار $T_d(s)^k z^{-s}$ حول $s=s_j$ تقریباً مانند رفتار $T_d(s)^k z^{-s}$ حول $s=\rho$ است [۱۰]، [۱۱].

با محاسبات ساده اما طولانی (تکرار رابطهٔ (۵)) این معادلهٔ به دست می‌آید:

$$P^{(k)}(x)=\sum_{\ell_1,\dots\ell_{d-1}}\binom{k-1}{\ell_1,\dots,\ell_{d-1}}P^{(1)}\left(p_1^{\ell_1}\dots p_{d-1}^{\ell_{d-1}}p_d^{k-1-\sum_{i=1}^{d-1}\ell_i}x\right)$$

که در آن $\ell_1+\dots+\ell_{d-1}=k-1$.

لم ۲ فرض کنید $x=re^{i\theta}$ که در آن $r\geq 0$ و $|\theta|\leq\pi$. در این‌صورت

$$|e^x P^{(k)}(x)|\leq e^r P^{(k)}(r)e^{-cr\theta^2}. \qquad (۹)$$

برهان چون $e^x P^{(1)}(x)$ یک سری توانی با ضرایب نامنفی است بنابراین

$$|e^x P^{(1)}(x)|\leq e^r P^{(1)}(r). \qquad (۱۰)$$

برای سهولت در محاسبات قرار می‌دهیم:

$$C(k,\ell)=\binom{k-1}{\ell_1,\dots,\ell_{d-1}}.$$

با استفاده از نابرابری $1-\cos\theta\geq 2\theta^2/\pi^2$ داریم:

$$\begin{aligned}|e^x P^{(k)}(x)|\leq\sum_{\ell_1,\dots,\ell_{d-1}}C(k,\ell)&\left|\exp\left\{x\left(1-p_1^{\ell_1}\dots p_{d-1}^{\ell_{d-1}}p_d^{k-1-\sum_{i=1}^{d-1}\ell_1}\right)\right\}\right|\\&\times\exp\left\{rp_1^{\ell_1}\dots p_{d-1}^{\ell_{d-1}-\sum_{i=1}^{d-1}\ell_i}\right\}\\&\times P^{(1)}\left(P_1^{\ell_1}\dots P_{d-1}^{\ell_{d-1}}P_d^{k-1-\sum_{i=1}^{d-1}\ell_i}r\right)\\=\sum_{\ell_1,\dots,\ell_{d-1}}C(k,\ell)&\left|\exp\left\{r\cos\theta\left(1-p_1^{\ell_1}\dots p_{d-1}^{\ell_{d-1}}p_d^{k-1-\sum_{i=1}^{d-1}\ell_i}\right)\right\}\right|\\&\times\exp\left\{rp_1^{\ell_1}\dots p_{d-1}^{\ell_{d-1}}p_d^{k-1-\sum_{i=1}^{d-1}\ell_i}\right\}\end{aligned}$$

$$\times P^{(1)}\left(P_1^{\ell_1}\dots P_{d-1}^{\ell_{d-1}}P_d^{k-1-\sum_{i=1}^{d-1}\ell_i}r\right)$$

$$=\sum_{\ell_1,\dots,\ell_{d-1}}\left|ex\left\{\left(1-2\theta^2/\pi^2\right)r\left(1-p_1^{\ell_1}\dots p_{d-1}^{\ell_{d-1}}p_d^{k-1-\sum_{i=1}^{d-1}\ell_i}\right)\right\}\right|$$

$$\times\exp\left\{rp_1^{\ell_1}\dots p_{d-1}^{\ell_{d-1}}p_d^{k-1-\sum_{i=1}^{d-1}\ell_i}\right\}$$

$$\times C(k,\ell)P^{(1)}\left(P_1^{\ell_1}\dots P_{d-1}^{\ell_{d-1}}P_d^{k-1-\sum_{i=1}^{d-1}\ell_i}r\right).$$

از طرفی

$$P_1^{\ell_1}\dots P_{d-1}^{\ell_{d-1}}P_d^{k-1-\sum_{i=1}^{d-1}\ell_i}<p_1^{k-1}<p_1.$$

بنا بر این

$$|e^xP^{(k)}(x)|\le e^{-2r\theta^2(1-p_1)/\pi^2}e^rP^{(r)}(x).$$

با انتخاب $c=2(1-p_1)/\pi^2>0$، برهان کامل می‌شود.

جهت برگشت از پواسن سازی، باید انتگرال کوشی

$$\mathrm{E}(B_{n.k})=\frac{n!}{2\pi i}\int_{|x|=n}e^xP^{(k)}(x)\frac{dx}{x^{n+1}}$$

$$=\frac{n!n^{-n}}{2\pi}\int_{|\theta|\le\pi}e^{nc^{i\theta}}P^{(k)}(ne^{i\theta})e^{-in\theta}d\theta$$

محاسبه شود [۱۵]. در زیر قضیهٔ اصلی را بیان و با روشی متفاوت با روشی که پارک و همکاران [۱۴] برای حالت $d=2$ به‌کار گرفتند، آن را ثابت می‌کنیم.

قضیه ۱. فرض کنید k و n اعدادی صحیح باشند و $\varepsilon>0$ مفروض باشد. در این‌صورت

آ) اگر

$$\left(-\log p_1(t)\right)^{-1}+\varepsilon\le\frac{k}{\log n}\le\frac{T_d(-2)}{-p_1(t)^2\log p_1(t)-p_2(t)^2\log p_2(t)}-\varepsilon,$$

آن‌گاه

$$\mathrm{E}(B_{n,k})=B\left(\rho_{n,k},\log_{\frac{p_1(t)}{p_2(t)}}p_1(s)^kn\right)\frac{T_d(\rho_{n,k})^kn^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}}\left(1+O\left(\frac{1}{\sqrt{k}}\right)\right),$$

که در آن

$$B(\rho,x)=\sum_{j\in z}f(\rho+it_j)\Gamma(\rho+it_j+1)e^{-2j\pi ix}$$

یک تابع دوره‌ای ناصفر با دورهٔ ۱ در x است.

ب) اگر $\frac{k}{\log n}\ge\frac{T_d(-2)}{-p_1(t)^2\log p_1(t)-p_2(t)^2\log p_2(t)}+\varepsilon$، آن‌گاه

$$\mathrm{E}(B_{n,k})=n^2\left(1-\sum_{i\ne j}^{d}p_j(1-p_i)\right)T_d(-2)^{k-1}(1+O(n^{-\eta})),$$

که در آن $\eta>0$.

پ) اگر $k/\log n$ نزدیک به $\frac{T_d(-2)}{-p_1(t)^2\log p_1(t)-p_2(t)^2\log p_2(t)}$، آن‌گاه $\xi=o((\log n)^{\frac{1}{6}})$،

$$\mathrm{E}(B_{n,k})=n^2\left(1-\sum_{i\neq j}^{d}p_j(1-p_i)\right)T_d(-2)^{k-1}\Phi(\xi)\left(1+O\left(\frac{1+|\xi|^3}{\sqrt{\log n}}\right)\right),$$

که در آن $\Phi(x)$ تابع توزیع نرمال استاندارد است.

برهان آ) فرض کنید $0<\theta<\pi/2$. در این‌صورت

$$\left|\frac{n!n^{-n}}{2\pi}\int_{\vartheta_0\le|\vartheta|\le\pi}e^{n\vartheta^{i\theta}}P^{(k)}(ne^{i\theta})e^{-in\vartheta}d\vartheta\right|$$
$$\le P^{(k)}(n)\frac{n!n^{-n}e^n}{2\pi}\int_{\vartheta\le|\vartheta|\le\pi}e^{-cn\vartheta^2}d\vartheta$$
$$=O\left(P^{(k)}(n)e^{-c\vartheta_0^2 2n}\right).$$

قرار دهید

$$f(s)=\frac{\left(\sum_{i\neq j}^{d}p_j(1-p_i)^{-s-1}-1\right)}{T_d(s)}.$$

بنابر نابرابری (۹)؛

$$\frac{n!n^{-n}}{2\pi}\int_{|\vartheta|\le\vartheta_0}e^{ne^{t,\vartheta}}P^{(k)}(ne^{i\theta})e^{-in\vartheta}d\vartheta$$
$$=\frac{n^{-\rho}T_d(\rho)^k}{\sqrt{2\pi\beta(\rho)k}}\sum_{|j|\le j_0}\Gamma(\rho+it_j)f(\rho+it_j)$$
$$\times\frac{n!n^{-n}}{2\pi}\int_{|\vartheta|\le\vartheta_0}e^{ne^{i\vartheta}}-^{in\vartheta}e^{i\vartheta(\rho+it_j)}d\vartheta.(1+O(k^{-1/2}))$$
$$=\frac{n^{-p}T_d(\rho)^k}{\sqrt{2\pi\beta(\rho)k}}\sum_{|j|\le j_0}\Gamma(\rho+it_j)f(\rho+it_j)$$
$$\times\frac{n!n^{-n}e^n}{2\pi}\int_{|\vartheta|\le\vartheta_0}e^{-\frac{1}{2}n\vartheta^2}.(1+O(n|\vartheta|^3)+O(|t_j\vartheta|)d\vartheta$$
$$\times\left(1+O(k^{-1/2})\right)$$
$$=\frac{n^{-\rho}T_d(\rho)^k}{\sqrt{2\pi\beta(\rho)k}}\sum_{|j|\le j_0}\Gamma(\rho+it_j)f(\rho+it_j)$$
$$\times\left(1+O(|t_j|n^{-1/2}\right)+O(k^{-1/2}))$$
$$=P^{(k)}(n)\left(1+O(k^{-1/2})\right).$$

حال نتیجهٔ مطلوب با استفاده از لم ۵ [۵] حاصل می‌شود.

ب) در دامنهٔ $\frac{k}{\log n}\ge\alpha_2+\varepsilon$، می‌توان روش مشابه‌ای را به‌کار گرفت. ابتدا مسیر انتگرال‌گیری را به $\Re(s)=\rho<-2$ انتقال می‌دهیم. چون تابع تحت انتگرال دارای یک تکینی قطبی در $s=-2$ است، بنابراین

$$P^{(k)}(x)=\left(1-\sum_{i\neq j}^{d}p_j(1-p_i)\right)x^2T_d(-2)^{k-1}$$
$$+\frac{1}{2}\int_{-\infty}^{\infty}\ x^{-s}\Gamma(\rho+it+1)\,f(\rho+it)T(\rho+it)^k\,dt.$$

با توجه به تعریف $T_d(s)$،

$$\frac{1-\sum_{i\neq j}^{d}p_j(1-p_i)}{T_d(-2)}<0.$$

بنابراین مجدداً بنا بر رابطهٔ (۹) داریم:

$$\mathrm{E}(B_{n,k})=n^2\left(1-\sum_{i\neq j}^{d}p_j(1-p_i)\right)T_d(-2)^{k-1}(1+O(n^{-\eta}),$$

که در آن $\eta>0$.

پ) حال فرض کنید $\xi=o\left((\log n)^{\frac{1}{6}}\right)$. در این صورت با انتقال خط انتگرال به نقطهٔ زینی

$$\Re(s)=\rho=-2-\frac{\xi}{\sqrt{\alpha_2\beta(-2)\log n}}+O\left(\frac{\xi^2}{\log n}\right).$$

داریم:

$$\mathrm{E}(B_{n,k})=n^2\left(1-\sum_{i\neq j}^{d}p_j(1-p_i)\right)T_d(-2)^{k-1}\Phi(\xi)\left(1+O\left(\frac{1+|\xi|^3}{\sqrt{\log n}}\right)\right).$$

نتیجه ۱. اگر در قضیهٔ ۱، $d=2$ آن‌گاه

$$1-\sum_{i\neq j}^{2}p_j(1-p_i)=1-p_1^2-p_2^2=2p_1p_2$$

که نتایج پارک و همکاران [قضیه‌های ۱الی ۱۴،۴] را نتیجه می‌دهد.

آماره‌ها

آ) عمق یک گره

فاصله از ریشه تا یک گره، عمق آن گره نامیده می‌شود. در این تعریف منظور از فاصله تعداد یال‌ها روی مسیر از ریشه تا گره است [۴]. فرض کنید این متغیر تصادفی با D_n نشان داده شود. در این‌صورت توزیع این متغیر تصادفی برابر تقسیم امید ریاضی نمایه افقی خارجی بر تعداد رشته‌ها است [۳]، یعنی:

$$P(D_n=k)=\frac{E(B_{n,k})}{n}.$$

قضیهٔ ۲. فرض کنید

$$H=p_1(t)\left(-\log p_1(t)\right)^2+p_2(t)\left(-\log p_2(t)\right)^2$$

و

$$h=p_1(t)\log\frac{1}{p_1(t)}+p_2(t)\log\frac{1}{p_2(t)}.$$

بنا براین برای هر n و k که در شرط $|k-\frac{1}{h}\log n|=o\left((\log n)^{2/3}\right)$ صدق می کنند،

$$P(D_n=k)=\frac{B\left(-1,\log_{\frac{p_1(t)}{p_2(t)}}p_1(s)^k n\right)}{\sqrt{2\pi(H-h^2)/h^3\log n}}\exp\left(-\frac{\left(k-\frac{1}{h}\log n\right)^2}{2(H-h^2)/h^3\log n}\right)$$
$$\times\left(1+O\left(\frac{1}{\sqrt{\log n}}+\frac{|k-\frac{1}{h}\log n|^3}{(\log n)^2}\right)\right).$$

برهان نتیجهٔ مستقیم قضیهٔ ۱ است.

ب) ارتفاع ترای

طول درازترین مسیر از ریشهٔ درخت، ارتفاع درخت نامیده می‌شود. با این تعریف اگر H_n، بیانگر ارتفاع ترای باشد، آن‌گاه $H_n=\max\left\{j;B_{n,j}>0\right\}$

قضیه ۳. فرض کنید $F_{H_n}(k)$ تابع توزیع احتمال H_n باشد. برای هر $k\geq 0$ وقتی $n\to\infty$،

$$F_{\mathrm{H}_n}(k)=\exp\left\{-\frac{1}{2}\exp\left(-\frac{4}{\log T_d(-2)}k+2\log n\right)\right\}+o(1).$$

برهان فرض کنید

$$G_k(x)=\sum_{n\geq 0}F_{\mathrm{H}_n}(k)\frac{x^n}{n!}$$

در این‌صورت دقیقاً مانند آن‌چه در مورد نمایه افقی خارجی برقرار بود، داریم.

$$G_k(x)=\prod_{i=1}^{d}G_{k-1}(p_i x),\qquad k\geq 2.$$

حال با روش مشابه با تحلیل نمایه، برهان کامل می‌شود (برای چگونگی انجام مراحل به [۲] و[۱۳] نگاه کنید).

نتیجه‌گیری

در این مقاله، از طریق یک رهیافت ترکیبیاتی خاص، میانگین نمایهٔ افقی خارجی در ترای محاسبه شد. نتایج از طریق ارتباط نمایه افقی خارجی با عمق گره و ارتفاع ترای به این پارامترها تسری داده شد. این رهیافت را می‌توان برای درخت‌های جستجوی رقومی با رشته‌های بیش از دو نوع رقم نیز به‌کار برد.

تقدیر و تشکر

از پیشنهادهای ارزندهٔ داوران گرامی و هیأت تحریریه محترم مجله که باعث بهبود مقاله شده است، تشکر و قدردانی می‌کنیم.

منابع

1. Briandais R. De La., "File searching using variable length keys, in Proceedings of the AFIPS Spring Joint Computer Conference", AFIPS Press, Reston, Va. (1959) 295-298.
2. Clement J., Flajolet P., Vallee B., "Dynamical sources in information theory: a general analysis of trie structures", Algorithmica, 29 (2001) 307-369.
3. Devroye L, "A study of trie-like structures under the density model", Annals of Applied Probaility, 2 (1992) 402-434.
4. Devroye L., "A note on the average depth of tries", Computing, 28 (1982) 367-371.
5. Drmota M., Szpankowski W., "The expected profile of digital search trees", Journal of Combinatorial Theory, Series A, 118 (2011) 1939-1965.
6. Flajolet P., Sedgewick R., "Analytic combinatorics", Cambridge University Press, Cambridge (2008).
7. Fredkin E., "Trie memory, Communications of the ACM", 3 (1960) 490-499.
8. Gusfield D. "Algorithms on strings, and sequences", Cambridge University Press, Cambridge (1997)..
9. Kazemi R., Vahidi-Asl M. Q., "The variance of the profile in digital search trees", Discrete Mathematics and Theoretical Computer Science, 13: 3 (2011) 21-38.
10. Kazemi R., Delaver S., "The moments of the profile in random binary digital trees", Journal of Mathematics and Computer Science, 6 (2013)176-190.
11. Kazemi R., Vahidi-Asl M. Q., "Probabilistic analysis of the asymmetric digital search trees", International Journal of Nonlinear Analysis and Applications, 6 (2) (2015) 161-173.
12. Kukich k., "Techniques for automatically correcting words in text", ACM Computing Surveys (1992)377-439.
13. Mahmoud H., "Evolution of random search tress", John Wiley and Soun Inc., New York (1992).
14. Park G., Hwang H. K., Nicodeme P., Szpankowski W., "Profile of tries, SIM J. Computing", 39 (2009) 1821-1880.
15. Szpankowski W., "Average case analysis of algorithms on sequences", Wiley, New York (2001).