В отговор на въпросите и критичните бележки на журито

Стефан Владимиров Герджиков

Настоящият текст коментира въпросите и основните критичните бележки в рецензиите и мненията на членовете на журито, които възникват във връзка с научните проблеми в предложения от мен дисертационен труд:

Ефективен алгоритъм за приближено търсене в регулярни множества

Той излага становището на автора му по конкретните въпроси в детайли, които не позволяват тяхното пълно устно представяне. Целта му е да послужи за евентуална основа на понататъшна дискусия по същество.

Този текст няма за цел да подменя нито целия, нито части от представения дисертационен труд. В него не се оспорва справедливостта на критиките на журито, а се възприема като конструктивна, като се уточнява как някои от нейните съществени точки могат да бъдат взети под внимание.

Нотационните и езиковите бележки на журито биха довели до безспорно опростяване на изложението на дисертационния труд и биха допринесли за неговото по-лесно четене и разбиране. Те, обаче, не са тема на настоящия текст.

В параграф 1 е отговорено на въпроса на академик Дренски, като изчерпателният теоретичен отговор е отложен в приложение A. В параграфи 2 и 3 са изложени коментарите на автора по отношение на основните критики за коректността на лема 7.3.2 и дефинициите от глава 8. Лемата е коригирана според препоръките на рецензента и е доказана в приложение B. В приложение C са дадени корекциите, необходими в дефинициите от глава 8.

1 Характеристиките памет и време на алгоритъма от глава 8.

Конкретните стойности в (мега)байти за индекса и в (мили)секунди за алгоритъма за търсене зависят от конкретната реализация и вътрешни характеристики на машината, върху която този алгоритъм се изпълнява. Експериментите от глава 8, изпълнени на компютър с 8-ядрен процесор 2.4 GHz Intel Xeon и 64 GB RAM, имат следните характеристики:

корпус (големина)	памет за индекс	време за трениране	време за търсене	средно време за търсене
TCD 1641	76.2 MB	22.14 sec	20.95 sec	1.13 msec.
IMPACT BG ($k = 1$)	255.6 MB	15.65sec	33.85 sec	1.03 msec.
IMPACT BG ($k = 2$)	259.5 MB	15.87sec	156.93 sec	5.63 msec.
IMPACT BG ($k = 3$)	262.3 MB	17.01sec	15.40 sec	0.67 msec.
IMPACT BG ($k = 4$)	264.8 MB	18.33sec	8.75 sec	0.49 msec.
IMPACT BG ($k = 5$)	267.4 MB	19.66sec	6.27 sec	0.48 msec.
IMPACT BG ($k = 6$)	270.2 MB	20.98sec	3.26 sec	0.41 msec.
ICAMET	863.4 MB	653.79 sec ($\sim$11 min)	93.30 sec ($\sim$1 min 30 sec)	1.53 msec
TREC-5	401 MB	1407.23 sec. ($\sim$22min)	2522.65 sec. ($\sim$42min)	0.24 msec.

От теоретична гледна точка, отговорът на въпроса на академик Дренски е изложен в приложение А.

Това, което трябва да се отбележи тук е следното. В експериментите върху българския корпус, IMPACT BG, резултатите са получени с основния алгоритъм, изложен в глава 8. Експериментите върху TCD 1641, TREC-5 и ICAMET предполагат наличието на неречникови думи. Затова при тях е използвана забележка 8.2.12. Допълнително, понеже в същите тези експерименти се налага обработката не само на единични думи, но и на фрази, използвана е и забележка 8.2.10 с праг $|Hypotheses| \leq 10000$.

2 Коректност на лема 7.3.2.

Критиката на професор Скордев относно споменатата лема е основателна и напълно справедлива. Както отбелязва рецензентът, лема 7.3.2 може да се коригира по следния начин:

Лема 1 *Нека $||.||$ е матрична норма, $t \in \mathbb{R}_+^{|\Sigma|}$, а $\mathcal{A}$ е краен автомат без ε-преходи, с n състояния и с матрица на съседство $M_{\mathcal{A}}(\boldsymbol{t})$. Тогава ако*

$$||M_{\mathcal{A}}(\boldsymbol{t})|| < 1.$$

то:

1. *за елементите на матриците $M_{\mathcal{A}}^N(\boldsymbol{t}) = (a_{j,k}^{(N)}(\boldsymbol{t}))_{j,k}$ е изпълнено, че редовете $\sum_{N=0}^{\infty} a_{j,k}^{(N)}(\boldsymbol{t})$ са сходящи за всеки $1 \leq j,k \leq n$ и $a_{j,k}^*(\boldsymbol{t}) = \sum_{N=0}^{\infty} a_{j,k}^{(N)}(\boldsymbol{t})$.*
2. *$\frac{\partial M_{\mathcal{A}}^*}{\partial t_i}(\boldsymbol{t})$ е добре дефинирана за всяко $i = 1 \ldots |\Sigma|$ и има елементи $\frac{\partial a_{j,k}^*}{\partial t_i}(\boldsymbol{t})$.*

Изменението на лемата е съществено и за нейното доказателство професор Скордев отбелязва, че *е необходимо да се положат някои допълнителни грижи.* Той посочва как това може да стане за $||.||_\infty$. В този параграф ще отбележим само, че предложената корекция лесно се обобщава за някои класически матрични норми. В приложение В ще уточним как доказателството на лема 1 може да бъде бъде извършено.

Идеята на професор Скордев в случая $||.|| = ||.||_\infty$ се основава на факта, че $||Av||_{L_1} \leq c||Av||_{L_\infty} \leq c||A||_\infty||v||_{L_\infty}$, където c е константа, независеща нито от $n \times n$-мерната матрицата A, нито от от n-мерния вектор v, а с L_1 и L_∞ са означени съответните евклидови норми на вектори.

Съществуването на такава константа се запазва и при по-общи предпоставки. Именно, достатъчно е матричната норма $||.|| : \mathcal{M}_{n,n}(\mathbb{R}) \to \mathbb{R}^+$ да притежава следното свойство: има реално число $p \geq 1$, че за всяка матрица $A \in \mathcal{M}_{n,n}(\mathbb{R})$ и всеки вектор $v \in \mathbb{R}^n$ е изпълнено:

$$||Av||_{L_p} \leq ||A||||v||_{L_p}, \text{ където}$$
$$||v||_{L_p} = \left(\sum_{i=1}^{n} |v_i|^p\right)^{\frac{1}{p}}.$$

Това свойство е изпълнено за всички, съгласувани норми $||.||_p$:

$$||A||_p = \sup_{v:||v||_{L_p}=1} ||Av||_{L_p},$$

а също и за нормата $||.||_\chi$, за която:

$$||A||_\chi = \max\{|\lambda| \,|\, det(A - \lambda E) = 0\},$$

където E е единичната матрица.

3 Коректност на дефинициите в глава 8

Забележките на професор Скордев във връзка с пропуските в основните дефиниции в тази глава са основателни. Те следва да бъдат коригирани както следва:

1. в дефиниции 8.1.2 предпоставката $V \in \mathcal{S}_\mathcal{W}$ следва да отпадне.

2. в дефиниция 8.1.11 следва да се уточни, че втората величина в минимума се разглежда само когато (под)дървета, които участват в нея са дефинирани.

3. в дефиниця 8.2.1 във втория случай знакът за принадлежност следва да бъде заменен със знака за непринадлежност. Вместо дефиниця 8.1.2 следва да се има предвид дефиницията с корекцията от точка 1.

Конкретно горните корекции означават, че съответните дефиниции следва да бъдат заменени с тези, посочени в приложение C.

За информация относно данните за корпуса TREC5 и публикациите от съответния форум следва да бъдат посочени следните два адреса:

http://trec.nist.gov/data/t5_confusion.html

http://trec.nist.gov/pubs/trec5/t5_proceedings.html.

Приложение A

Паметта, необходима за представянето на индекса и операциите, използвани от алгоритъма теоретично може да бъде оценена като:

$$O\left(\sum_{U\in\mathcal{D}} |U| + \sum_{V\in\mathcal{N}} |V| + \sum_{(U,V)\in\mathcal{I}} |U|^2|V|\right)$$

Наистина, от предварителните резултати в Глава 2 следва, че паметта за структурите $\mathcal{S}_\mathcal{N}$ и $\mathcal{S}_\mathcal{D}$ е линейна относно броя символи в $\mathcal{N}$ и $\mathcal{D}$ съответно. Допълнителната памет, необходима за представянето на автоматично извлечените операции Op, може да се оцени въз основа на разсъжденията от доказателството на леми 8.1.13 и 8.1.17. Така една оценка отгоре за броя на операциите Op е:

$$|Op| \in O\left(\sum_{(U,V)\in\mathcal{I}} |U|^2|V|\right).$$

На практика, в случаите, които са представени в параграф 8.3 този брой е много по-малък, именно $|Op| \in O(|\mathcal{S}_\mathcal{N}|)$.

Времевата сложност на алгоритъма за търсене е изведена в твърдение 8.2.9:

$$O(|V| + \sum_{h\in Hypotheses} (|U(h)| + \log|Hypotheses|)),$$

където V е дадената в заявката дума, множеството $Hypotheses$ са генерираните от алгоритъма хипотези. Неявният и неизвестният предварително параметър $Hypotheses$ е този, който определя времевата сложност на практика. Както е обяснено в забележки 8.2.10 и 8.2.11 големината на този параметър може да бъде контролирана.

Доколкото при стойностина този параметър от порядъка на 10000 допълнителната памет, която изисква алгоритъмът за търсене в интересните за практиката случаи се доминира от големината на индекса, то тя може да бъде пренебрегната.

Приложение B

Доказателство: (на лема 1)

1. Нека $B^{(K)}(\boldsymbol{t}) = \sum_{N=0}^{K} M_\mathcal{A}^N(\boldsymbol{t})$ с елементи $b_{j,k}^{(K)}(\boldsymbol{t})$ за $1 \leq j,k \leq n$. Тъй като векторът $\boldsymbol{t}$ е покоординатно неотрицателен, то и елементите на матрицата $M_\mathcal{A}(\boldsymbol{t})$ са неотрицателни. Следователно за всеки два фиксирани индекса j,k редицата $b_{j,k}^{(K)}(\boldsymbol{t})$ е ненамаляваща по $K \in \mathbb{N}$ и в частност нейните члинове са неотрицателни. Освен това е ясно, че:
$$b_{j,k}^{(K)}(\boldsymbol{t}) = \sum_{N=0}^{K} a_{j,k}^{(N)}(\boldsymbol{t}).$$

За дадено $\boldsymbol{t} \in \mathbb{R}_+^{|\Sigma|}$ и $K \in \mathbb{N}$ да означим с $\beta(K)$:

$$\beta(K) = \max_{1 \leq j,k \leq n} b_{j,k}^{(K)}(\boldsymbol{t}),$$

а $(\iota(K), \kappa(K))$ да бъде конкретна двойка от индекси, за които е в сила равенството:

$$\beta(K) = b_{\iota(K),\kappa(K)}^{(K)}(\boldsymbol{t}).$$

Тъй като има краен брой различни двойки индекси (j,k), за които $1 \leq j,k \leq n$, а редицата от двойки индекси $\{(\iota(K), \kappa(K))\}_{K=0}^{\infty}$ е безкрайна, то съществува двойка от индекси (j_0, k_0) и безкрайна редица от естествени числа $K_0 < K_1 < \cdots < K_s < \ldots$, за които:

$$\forall s \in \mathbb{N}[(j_0, k_0) = (\iota(K_s), \kappa(K_s))].$$

Сега за всяка двойка индекси (j,k) и за всяко естествено число s е изпълнено, че:

$$0 \leq b_{j,k}^{(K_s)}(\boldsymbol{t}) \leq \beta(K_s) = b_{j_0,k_0}^{(K_s)}(\boldsymbol{t}),$$

където първото неравенство следва от неотрицателността на елементите на матрицата $B^{(K_s)}(\boldsymbol{t})$, а второто следва от максималността на $\beta(K_s)$. Нещо повече, тъй като $b_{j,j}^{(0)} = 1$ за всяко $j \leq n$, то $\beta(K_s) > 0$ за всяко s. От горните неравенства следва, че за всеки $1 \leq j,k \leq n$ редицата:

$$\frac{b_{j,k}^{(K_s)}(\boldsymbol{t})}{\beta(K_s)}$$

е коректно дефинирана и ограничена между 0 и 1. Следователно съдържа сходяща подредица. Тъй като броят на редиците от разглеждания вид е краен, то може да намерим безкрайна подредица $K_0' < K_1' < \ldots K_s' < \ldots$ на редицата $\{K_s\}_{s=0}^{\infty}$ със свойството, че:

$$\forall j,k \leq n \left[\frac{b_{j,k}^{(K_s')}(\boldsymbol{t})}{\beta(K_s')} \text{ е сходяща.} \right]$$

Полагаме $C^{(s)}$ да бъде матрицата $C^{(s)} = \frac{1}{\beta(K_s')} B^{(K_s')}(\boldsymbol{t})$. Нека C^{∞} е матрицата с елементи $c_{j,k}^{\infty}$, които се определят като:

$$c_{j,k}^{\infty} = \lim_{s \to \infty} \frac{b_{j,k}^{(K_s')}(\boldsymbol{t})}{\beta(K_s')}.$$

Тогава лесно се вижда, че $\lim_{s\to\infty}||C^{(s)}|| = ||C^{\infty}||$. Наистина ако положим $E_{j,k}$ да бъдат матриците с 1 на позиция (j,k) и 0 на всички останали позиции, и използваме неравенството на триъгълника за матричната норма, получаваме:

$$|||C^s||-||C^{\infty}||| \leq ||C^s-C^{\infty}|| = ||\sum_{j,k=1}^{n,n}(c_{j,k}^{(s)}-c_{j,k}^{\infty})E_{j,k}|| \leq \sum_{j,k=1}^{n,n}|c_{j,k}^{(s)}-c_{j,k}^{\infty}|||E_{j,k}||.$$

Тъй като $||E_{j,k}||$ е крайно число и $c_{j,k}^{\infty}$ е границата на $c_{j,k}^{(s)}$ при s, клонящо към безкрайност, то дясната страна на неравенството клони към 0 при s, клонящо към безкрайност, което показва, че и абсолютната стойност на разликата от нормите на $C^{(s)}$ и C^{∞} също е 0. Оттук следва, че $\lim_{s\to\infty}||C^{(s)}|| = ||C^{\infty}||$. Да отбележим още, че елементът $c_{j_0,k_0}^{(s)} = 1$ за всяко s и следователно, $c_{j_0,k_0}^{\infty} = 1$. Това показва, че C^{∞} не е нулева и следователно $||C^{\infty}|| \neq 0$.

Сега може да покажем, че редицата $\beta(K_s')$ е ограничена. Наистина имаме, че:

$$\beta(K_s')||C^{(s)}|| = ||B^{(K_s')}(\boldsymbol{t})|| = ||\sum_{N=0}^{K_s'}M_{\mathcal{A}}(\boldsymbol{t})^N|| \leq \sum_{N=0}^{K_s'}||M_{\mathcal{A}}(\boldsymbol{t})||^N \leq \frac{1}{1-||M_{\mathcal{A}}(\boldsymbol{t})||},$$

където последното неравенство следва от условието, че $||M_{\mathcal{A}}(\boldsymbol{t})|| < 1$, а предпоследното от неравенството на триъгълника за матрични норми и от неравенството $||AB|| \leq ||A||||B||$.

Сега от факта, че $\lim_{s\to\infty}||C^{(s)}|| = ||C^{\infty}||$ и тъй като $||C^{\infty}|| \neq 0$, следва, че за достатъчно големи стойности на s $||C^{(s)}|| \geq \frac{1}{2}||C^{\infty}|| > 0$. Оттук получаваме, че за достатъчно големи s е в сила и неравенството:

$$\beta(K_s') \leq \frac{1}{1-||M_{\mathcal{A}}(\boldsymbol{t})||}\frac{1}{||C^{(s)}||} \leq \frac{1}{1-||M_{\mathcal{A}}(\boldsymbol{t})||}\frac{2}{||C^{\infty}||}.$$

Тъй като $\beta(K_s')$ е монотонна по s и е ограничена, то тя е сходяща. Тъй като $\beta(K_s') \geq b_{j,k}^{(K_s')}(\boldsymbol{t})$ за всеки $1 \leq j,k \leq n$, то и редиците $b_{j,k}^{(K_s')}(\boldsymbol{t})$ са ограничени и от тяхната монотонност следва, че са сходящи.

Накрая от това, че всяка от редиците $b_{j,k}^{(K)}(\boldsymbol{t})$ е монотонна и съдържа сходяща подредица, следва, че и самите редици $b_{j,k}^{(K)}(\boldsymbol{t})$ са сходящи.

С това доказахме, че редовете $\sum_{N=0}^{\infty}a_{j,k}^{(N)}(\boldsymbol{t})$ са (абсолютно) сходящи при условие, че $||M_{\mathcal{A}}(\boldsymbol{t})|| < 1$ и следователно матрицата $M_{\mathcal{A}}^*(\boldsymbol{t})$ е добре дефинирана.

2. Тъй[1] като елемнтите на $M_{\mathcal{A}}(\boldsymbol{t})$ са непрекъснати (линейни) функции на t, то в достатъчно малка околност на t също ще бъде изпълнено, че $||M_{\mathcal{A}}(\boldsymbol{t'})|| < 1$. Следователно матриците $M^*_{\mathcal{A}}(\boldsymbol{t'})$ ще бъдат добре дефинирани и в околност на t. Също така, не е трудно да се види, че:

$$\sum_{N=0}^{K}\left(M^N_{\mathcal{A}}(\boldsymbol{t}) - M^N_{\mathcal{A}}(\boldsymbol{t'})\right) = \sum_{N=1}^{K}\sum_{N_1=0}^{N-1} M^{N_1}_{\mathcal{A}}(\boldsymbol{t})(M_{\mathcal{A}}(\boldsymbol{t}) - M_{\mathcal{A}}(\boldsymbol{t'}))M^{N-N_1-1}_{\mathcal{A}}(\boldsymbol{t'}).$$

Оттук директно следва, че:

$$\begin{aligned}\left|\left|\sum_{N=0}^{K}\left(M^N_{\mathcal{A}}(\boldsymbol{t}) - M^N_{\mathcal{A}}(\boldsymbol{t'})\right)\right|\right| &\leq ||M^*_{\mathcal{A}}(\boldsymbol{t})||||M_{\mathcal{A}}(\boldsymbol{t}) - M_{\mathcal{A}}(\boldsymbol{t'})||||M^*_{\mathcal{A}}(\boldsymbol{t'})|| \\ &\leq \frac{||M_{\mathcal{A}}(\boldsymbol{t}) - M_{\mathcal{A}}(\boldsymbol{t'})||}{(1-||M_{\mathcal{A}}(\boldsymbol{t})||)(1-||M_{\mathcal{A}}(\boldsymbol{t'})||)}.\end{aligned}$$

Тъй като елементите на $M_{\mathcal{A}}(\boldsymbol{t'})$ са непрекъснати функции на аргумента $\boldsymbol{t'}$, то получаваме, че $M^*_{\mathcal{A}}(\boldsymbol{t'})$ е непрекъсната в околност на $\boldsymbol{t}$. Сега не е трудно да се види, че:

$$\frac{M^*_{\mathcal{A}}(\boldsymbol{t}) - M^*_{\mathcal{A}}(\boldsymbol{t'})}{t_i - t'_i} = M^*_{\mathcal{A}}(\boldsymbol{t})\frac{M_{\mathcal{A}}(\boldsymbol{t}) - M_{\mathcal{A}}(\boldsymbol{t'})}{t_i - t'_i}M^*_{\mathcal{A}}(\boldsymbol{t'})$$

и тъй като при $\boldsymbol{t'} \to \boldsymbol{t}$ $(t'_i \neq t_i)$ имаме, че:

$$\frac{M_{\mathcal{A}}(\boldsymbol{t}) - M_{\mathcal{A}}(\boldsymbol{t'})}{t_i - t'_i} \to \frac{\partial M_{\mathcal{A}}}{\partial t_i}(\boldsymbol{t}),$$

то получаваме, че:

$$\frac{M^*_{\mathcal{A}}(\boldsymbol{t}) - M^*_{\mathcal{A}}(\boldsymbol{t'})}{t_i - t'_i} \to M^*_{\mathcal{A}}(\boldsymbol{t})\frac{\partial M_{\mathcal{A}}}{\partial t_i}(\boldsymbol{t})M^*_{\mathcal{A}}(\boldsymbol{t})$$

когато $\boldsymbol{t'}$ клони към $\boldsymbol{t}$ със стойности $t'_i \neq t_i$. Това показва, че частната производна на $M^*_{\mathcal{A}}$ в $\boldsymbol{t}$ по направление t_i съществува.

[1] Неявно, в тази част на доказателството използваме, че сходимостта, установена по-горе е равномерна. Това следва при предположение, че елементите на $M_{\mathcal{A}}(.)$ са непрекъснати в околност на вектора $\boldsymbol{t}$.

Приложение C

Definition 1 *(8.1.2) Given a finite set of words, $\mathcal{W}$, and* ***a word*** *$V \in \Sigma^*$, the longest proper suffix of V, $lps_{\mathcal{W}}(V) = lps(V)$ is the longest infix $V_1 \neq V$ s.t.:*

$$V_1 \in Suf(V) \text{ and } V_1 \in \mathcal{S}_{\mathcal{W}}.$$

Definition 2 *(8.1.11) Let $\mathcal{T}_{\mathcal{N}}(V)$ be a canonical tree for the distinx V in the set of noisy words,n $\mathcal{N}$. Let $\mathcal{CT}_{\mathcal{D}}(U)$ be a candidate tree for the distinx U in the dictionary, $\mathcal{D}$. We set $id = 1$, if U and V share a common last character, and $id = 0$, otherwise. Then the edit-distance between the trees $\mathcal{T}_{\mathcal{N}}(V)$ and $\mathcal{CT}_{\mathcal{D}}(U)$ is defined as follows:*

1. *if some of the trees $\mathcal{CT}_{\mathcal{D}}(U)$ or $\mathcal{T}_{\mathcal{N}}(V)$ has no subtrees, then:*

$$d_T(\mathcal{CT}_{\mathcal{D}}(U), \mathcal{T}_{\mathcal{N}}(V)) = \max(|U|, |V|) - id$$

2. *if $\mathcal{CT}_{\mathcal{D}}(U)$ and $\mathcal{T}_{\mathcal{D}}(V)$ both have left and right subtrees denoted with superspript (l) and (r), respectively, then:*

$$d_T(\mathcal{CT}_{\mathcal{D}}(U), \mathcal{T}_{\mathcal{N}}(V)) = \min \begin{cases} \max(|U|, |V|) - id \\ d_T(\mathcal{CT}^{(l)}_{\mathcal{D}}(U), \mathcal{T}^{(l)}_{\mathcal{N}}(V)) + d_T(\mathcal{CT}^{(r)}_{\mathcal{D}}(U), \mathcal{T}^{(r)}_{\mathcal{N}}(V)) \end{cases}$$

Definition 3 *(8.2.1) Let V be a word and $\mathcal{W}$ be a finite (nonempty) set of words. An approximate canonical tree $\widetilde{\mathcal{T}}_{\mathcal{W}}(V)$ is defined recursively as follows:*

1. *if $V \in \mathcal{S}_{\mathcal{W}}$, then $\widetilde{\mathcal{T}}_{\mathcal{W}}(V) = \mathcal{T}_{\mathcal{W}}(V)$.*
2. *if $V \notin \mathcal{S}_{\mathcal{W}}$, then $\widetilde{\mathcal{T}}_{\mathcal{W}}(V)$ is a tree with root V, right subtree $\mathcal{T}_{\mathcal{W}}(V_2)$ and left subtree $\widetilde{\mathcal{T}}_{\mathcal{W}}(V_1)$ where:*

$$V_2 = lps_{\mathcal{W}}(V) \text{ and } V = V_1 \circ V_2.$$