РЕЦЕНЗИЯ

от проф. дмн Димитър Скордев
на дисертацията на асист. Стефан Владимиров Герджиков
„EFFICIENT ALGORITHM
FOR THE APPROXIMATE SEARCH PROBLEM IN REGULAR SETS",
представена за присъждане на образователната и научна степен „доктор"
в професионално направление 4.5 „Математика"
по научната специалност 01.01.01 „Математическа логика"

Описание на представените материали

Както личи от нейното заглавие, дисертацията е на английски език. Основният ѝ текст е изложен на 164 страници и се състои от осем глави. Предхожда се от 12 страници увод, а се следва от пет страници добавка, пет страници заключение и пет страници библиография с данни за 67 източника. Към дисертацията са приложени следните материали на български език: автореферат (25 стр.), резюме (1 стр.), текст, озаглавен „Резюме на получените резултати и декларация за оригиналност на труда" (7 стр.) и автобиография на дисертанта (2 стр.). Приложени са още и електронни копия на една самостоятелна и три съвместни негови публикации по тематиката на дисертацията, а също електронни копия на следните документи: заявление на дисертанта до ръководителя на катедрата по математическа логика и приложенията ѝ, заповед за зачисляване, заповед за отчисляване (с право на защита), диплома за висше образование и приложение към нея, удостоверение за положени изпити в съответствие с индивидуалния учебен план на докторанта. Споменатата по-горе самостоятелна публикация (8 стр.) е отпечатана в „Доклади на БАН" през 2012 г., две от съвместните (4 стр. и 5 стр.) са в трудовете на международни конференции, проведени съответно през март и август 2013 г., а третата (36 стр.) се намира на сайта arXiv от януари 2013 г. и представлява пълна версия на първата от другите две съвместни публикации. В първата и третата от съвместните публикации съавтори на дисертанта са научният му ръководител доц. д-р Стоян Михов заедно с Петър Митанкин и Клаус Шулц, а във втората са Стоян Михов и Владислав Ненчев. От сведенията, дадени на страници 173-174 от дисертацията и страници 21-22 от автореферата, се вижда, че участието на дисертанта в съвместните публикации е съществено. Не открих сведения някоя от представените публикации да е била цитирана от друг автор.

Предмет на дисертацията

Когато са дадени една азбука Σ и едно крайно множество от двойки различни думи над Σ, като двойките се тълкуват като възможни замени на думи над Σ с други такива, заслужава внимание понятие за *разстояние* $d(U,V)$ от дума U до дума V, дефинирано като минималния брой незастъпващи се части на U, такива, че чрез техни замени измежду онези в множеството да може да се получи V (ако не е възможно V да се получи от U по този начин, считаме, че $d(U,V) = \infty$).[1] По-общо, бихме могли да разглеждаме случая, когато на всяка замяна е съпоставено положително число, наречено нейна *цена*, и вместо броя на заменяните части се разглежда сборът от цените на замените. В този случай, включвайки в множеството също замяната на коя да е еднобуквена дума със самата нея и приписвайки на тази замяна цена 0, можем при определянето на $d(U,V)$, вместо да разглеждаме произволни системи от незастъпващи се части на U, да се ограничим с онези, на които конкатенацията е U. Произволна наредена двойка (Op, c) от такова множество Op от замени и негов „ценоразпис“ c е наречена *ортографско разстояние*.[2] Крайна редица от елементи на Op се нарича *подравняване*. Ако конкатенацията на първите компоненти на членовете на едно подравняване ω е U, а конкатенацията на вторите е V, то ω се нарича подравняване на U с V, а думите U и V се наричат съответно лява и дясна проекция на ω. *Цена* на едно подравняване е сборът от цените на неговите членове. При тази терминология $d(U,V)$ е най-малката от цените на подравняванията на U с V (при уславяне, че най-малкият елемент на празно множество от естествени числа е ∞).

Основната задача, която се разглежда в дисертацията, е намирането на ефективен алгоритъм, който при дадено ортографско разстояние (Op, c), даден регулярен език над Σ и дадено рационално число q между 0 и 1 за всяка дума V над Σ да дава всички думи U от дадения език, удовлетворяващи неравенството $d(U,V) \leq q|V|$ (където $|V|$ е дължината на V). Важно поле за приложение на такива алгоритми е търсенето на кандидати за корекция на думи, в които са налице грешки, причинени от шум. Предишни резултати в това направление са били получени от други автори като например Юджин Майърс, предложил ефективен алгоритъм за случая, когато езикът се състои от поддумите на дадена дума, а разстоянието е Левенщейновото. В дисертацията методът на Майърс се усъвършенства и се обобщава за случая на езика $Inf(\mathcal{L})$ на поддумите на думите от произволен даден регулярен език $\mathcal{L}$ и за произволно ортографско разстояние. Намират се условия, при които предложеният общ метод има линейна оценка на средната времева сложност.

Друга задача, която се разглежда (в последната глава на дисертацията), е търсенето на подходящ списък от замени и подходящи техни вероятности въз основа на даден списък от примери за коригиране на сгрешена дума и даден речник на езика, както и приложението на това за автоматично

[1] Класически пример е разстоянието, разгледано от Владимир Левенщейн през 1965 г. – то съответства на множеството от следните три вида замени: на празната дума с коя да е еднобуквена, на произволна еднобуквена с празната и на коя да е еднобуквена с произволна друга еднобуквена. Разбира се в този случай $d(U,V)$ никога не е ∞.

[2] Елементите на Op се наричат операции при терминологията на дисертацията. Терминът „ортографско разстояние“ се използва в автореферата. В самата дисертация за същото понятие е използван терминът „edit-distance“.

коригиране на грешките.

Приноси в дисертацията

Ще направя преглед на по-важните резултати на автора на дисертацията, които са изложени в нея. Следвайки реда, в който са изброени те на страници 171-172 от дисертацията и страница 20 от автореферата, ще привеждам (в курсив) дадените на последната формулировки и ще правя своите коментари и преценки.

*1. **Общ апарат за изучаване на свойствата на подравнявания от думи и ефективното генериране на кандидати за корекция.** Този апарат, основаващ се на концепцията за множествата от подравнявания и списъци от разстояния, е разработен в Глава 4.*

Част от разглежданията в тази глава са за случая $\rho = 1$, където ρ е най-голямата от дължините на вторите компоненти в двойките от Op. В този случай се показва (в лема 4.1.1 и следствие 4.1.2), че когато една дума V над Σ е конкатенация на две думи V_0 и V_1, всяко подравняване ω с дясна проекция V е конкатенация на някое подравняване ω_0 с дясна проекция V_0 и някое подравняване ω_1 с дясна проекция V_1, и е ясно, че ако цената на ω не надминава сбора на две числа b_0 и b_1, то цената на ω_0 не надминава b_0 или цената на ω_1 не надминава b_1. Грубо казано, алгоритъмът, предложен в следващата глава, се основава освен на други рекурсивни зависимости още на използване на горното, докато е уместно това, при $|V_0| - |V_1| \in \{0,1\}$, $b_0 = q|V_0|$, $b_1 = q|V_1|$ и на обстоятелството, че при $q|V| < 1$ неравенството $d(U,V) \leq q|V|$ се удовлетворява само при $U = V$. Като подготовка за осъществяването на така описаната идея се доказват няколко твърдения в рекурсивен дух за множества от подравнявания с отнапред дадена дясна проекция и за частични функции от Σ^* към $\mathbb{N}$, които в определен смисъл представят тези множества при дадения език $\mathcal{L}$.

*2. **Нов алгоритъм за приближено търсене в произволно регулярно множество от думи.** Това е алгоритъмът по схемата разделяй и владей, който е представен подробно в Глава 5.*

В тази глава се предполага, че е дадено ортографско разстояние, удовлетворяващо условието $\rho = 1$. Алгоритъмът, който се предлага, при дадена дума V използва изображение $\alpha \mapsto V_\alpha$ на едно дървовидно крайно подмножество на $\{0,1\}^*$ в множеството на поддумите на V. А именно, полага се $V_\varepsilon = V$, а винаги, когато V_α е дефинирано и $q|V_\alpha| \geq 1$, $V_{\alpha 0}$ и $V_{\alpha 1}$ също са дефинирани, като конкатенцията им е V_α, а разликата $|V_{\alpha 0}| - |V_{\alpha 1}|$ е 0 или 1. За всяко α, за което V_α е дефинирано, се търсят онези думи U от $Inf(\mathcal{L})$, за които съществува подравняване на U с V_α с цена, ненадминаваща $q|V_\alpha|$; при това за всяка такава дума U се търси и най-малката цена на подравняване с посочените свойства. В случай, че $q|V_\alpha| < 1$, това е много лесно: ако $V_\alpha \in Inf(\mathcal{L})$, то единствената такава дума U е V_α и съществува единствено подравняване с въпросните свойства, като цената му е 0, а ако $V_\alpha \notin Inf(\mathcal{L})$, то няма дума U от $Inf(\mathcal{L})$, за която да съществува подравняване с тези свойства. С преодоляване на ред трудности, ръководен от много добра интуиция, авторът обаче е стигнал до удобен от изчислителна гледна точка начин за решаване на така описаната задача за кое да е α, за което $q|V_\alpha| \geq 1$, в случай, че задачата е вече решена за $\alpha 0$ и $\alpha 1$. С нужния брой прилагания на този начин се стига в крайна сметка до решение на задачата при $\alpha = \varepsilon$ и с това се намират всички думи U от $Inf(\mathcal{L})$, за които съществува подравняване на U с V с цена, ненадминаваща $q|V|$.

*3. **Нов алгоритъм за приближено търсене за произволно ортографско разстояние.** Това е алгоритъмът представен в Глава 6, който използва схемата разделяй и владей и решава проблема за приближено търсене за произволно ортографско разстояние.*

В глава 6 на дисертацията алгоритъмът от глава 5 е модифициран така, че да може да се използва и без да е налице равенството $\rho = 1$ (тъй като случаят $\rho = 0$ е прост, може да се предполага, че $\rho \geq 1$). Липсата на равенството $\rho = 1$ води до технически усложнения – например в общия случай не е вярно заключението на лема 4.1.1 и става нужда да се използва по-сложното твърдение на лема 6.1.1. Там, където в дефиницията на изображението $\alpha \mapsto V_\alpha$ по-рано фигурираше неравенството $q|V_\alpha| \geq 1$, сега заедно с него фигурира и неравенството $|V_\alpha| \geq 4(\rho - 1)$ и освен това става нужда да се работи не само с думите V_α, $V_{\alpha 0}$ и $V_{\alpha 1}$, но и с подходящи техни поддуми. За онези α, за които $q|V_\alpha| \geq 1$, но $|V_\alpha| < 4(\rho - 1)$, намирането на съответните думи U от $Inf(\mathcal{L})$ изисква повече труд и те могат да не са единствени. В крайна сметка обаче всички затруднения са преодолени успешно.

*4. **Практическа ефективност на предложените алгоритми.** Предложеният алгоритъм има свойството, че постепенно увеличава допустимото разстояние. По този начин броят на генерираните кандидати се поддържа разумно малък. Допълнителен аргумент за практическата ефективност на алгоритъма са и резултатите от Твърдения 1 и 2, които казват, че всеки кандидат за корекция в този алгоритъм се генерира веднъж, (по-точно този брой е O(1) – ограничен от глобална константа).*[3]

Съгласен съм с казаното в първите две изречения. Не мога да преценя обаче дали наистина могат да се разглеждат като аргумент за практическата ефективност на алгоритъма споменатите твърдения 1 и 2.

*5. **Теоретична ефективност на предложения алгоритъм.** Предложен е вероятностен подход, който теоретично аргументира ефективността на алгоритъма, Твърдение 3.*[4]

При доказателството на твърдение 3 е проявена голяма изобретателност. Фактически обаче твърдението само по себе си не аргументира ефективността на алгоритъма, защото изразът в скобите в дясната страна на неравенството не винаги е с крайна стойност. От значение е са и резултатите от следващия параграф – от тях става ясно, че при известни допълнителни предположения стойността на въпросния израз е крайна (формулировката и доказателството на съществената за целта лема 7.3.2 имат дефекти, но за

[3] Твърдения 1 и 2 са съответно твърдения 6.2.8 и 6.2.9 от дисертацията.

[4] Твърдение 3 е твърдение 7.2.2 от дисертацията.

щастие те са лесно отстраними[5]).

6. Нов подход за дефиниране на близост между думи. Глава 8 представя нов подход за дефиниране на близост между думи, който отчита цялостната структура на езика, а също и контекстната информация в рамките на самите думи. Концептуалните предимства на този подход са потвърдени експериментално, а ефективността на разработения алгоритъм е оценена в Твърдение 4.[6]

Идеята на споменатия подход изглежда плодотворна и данните за експериментално потвърждение на неговите предимства правят много благоприятно впечатление. За съжаление изложението на математическата същност на метода има дефекти, които ми попречиха да вникна подробно в нея (и в частност в споменатото твърдение). Ето главните от тях:

- В примера, илюстриращ разглежданията в подпараграф 8.1.1, фактически не се съобщава нищо конкретно за множествата $\mathcal{N}$ и $\mathcal{D}$ освен това, че думата *knoweth* принадлежи на $\mathcal{N}$, а думата *knows* принадлежи на $\mathcal{D}$ – по тази причина префиксите на първата са единствените думи, за които читателят може да провери дали удовлетворяват условията от дефиниция 8.1.1 при $\mathcal{W} = \mathcal{N}$ и аналогично е положението с удовлетворяването на въпросните условия при $\mathcal{W} = \mathcal{D}$.
- От дефиниции 8.1.2, 8.1.3 и 8.1.5 следва, че ако V е еднобуквена дума от $S_{\mathcal{W}}$, то дървото $\mathcal{T}_{\mathcal{W}}$ освен корен има и два листа. Фиг. 8.1 обаче не е в съгласие с това.
- В рекурсивната дефиниция 8.1.11 е пропуснат случаят, когато някое от двете дървета е тривиално, и поради това тя не дефинира нищо. В дясната ѝ страна е използвано означението за израз, включващ разглеждане на случаи, а такова в случая липсва.
- В дефиниция 8.2.1 има грешка – така, както е формулирана, и двете точки в нея се отнасят за случая, когато $V \in S_{\mathcal{W}}$. Явно някоя от тях е трябвало да бъде за противоположния случай, но възниква проблемът, че

[5]Тъй като компонентите на матрицата $M^*_{\mathcal{A}}(t)$ са безкрайни редове, няма как в общия случай да се говори за нейна норма, а когато може да се говори, няма как нормата да е безкрайна (според дефиниция 1.7.1 нормите на матрици са изображения на множеството на квадратните матрици с дадена размерност, които са с реални компоненти, в множеството на реалните числа). Поради тази причина формулировката на лема 7.3.2 не е коректна. Коректната формулировка би била, че при направените предположения степенните редове, които са компоненти на матриците $M^*_{\mathcal{A}}(t)$ и $\dfrac{\partial M^*_{\mathcal{A}}(t)}{\partial t_j}$, $j = 1, \ldots, |\Sigma|$, са сходящи. Разбира се, доказателството на лемата също се нуждае от изменения. За произволна норма би се наложило да се положат някои допълнителни грижи, но при нормата, спомената в края на параграф 1.7, коригираното по горния начин заключение ще може да се получи с малка промяна в доказателството, а именно като се докаже, че за всяко естествено число K са в сила неравенствата

$$\left\|\sum_{N=0}^{K} M_{\mathcal{A}}^{N}(t)\right\| \leq \frac{1}{1-\|M_{\mathcal{A}}(t\|}, \quad \left\|\sum_{N=0}^{K} \frac{\partial M_{\mathcal{A}}^{N}(t)}{\partial t_j}\right\| \leq \left\|\frac{\partial M_{\mathcal{A}}(t)}{\partial t_j}\right\| \frac{1}{(1-\|M_{\mathcal{A}}(t)\|)^2},$$

и с тяхна помощ се покаже, че множествата на частичните суми на споменатите редове са ограничени отгоре.

[6]Твърдение 4 е твърдение 8.2.9 от дисертацията.

в първата се използва означението $T_W(V)$, а във втората – означението $lps_W(V)$, като и двете означения са въведени в съответните дефиниции 8.1.5 и 8.1.2 при предположение, че $V \in S_W$.

Освен горните критични бележки имам и многобройни други, които се отнасят главно до изложението в други места от дисертацията (някои негови недостатъци също ми създадоха доста затруднения при проучването ѝ, а други, макар и по-дребни, водеха до излишно натоварване при четенето). Въпреки това равносметката ми за резултатите в нея е положителна. По-конкретно, смятам, че резултатите, изложени в глави 5, 6 и 7, са много ценни, както в теоретично, така и в приложно отношение, а авторът им е навлязъл дълбоко в тематиката и е преодолял сериозни трудности при тяхното получаване (относно ценността на резултатите от глава 8 в теоретично отношение бих могъл да се произнеса едва след запознаване с коректно изложение на математическата същност на метода).

Критични бележки

Някои критични бележки вече бяха направени. За да не удължавам прекомерно основния текст на рецензията, ще изложа повечето от останалите в едно доста обемисто приложение към нея, а тук ще се ограничa само с мнението, че изложението в глави 4 и 5 би спечелило, ако се направят следните изменения:

1. *Разграничаване в по-голяма степен на излагането на математическите свойства на използваните функции от описанието на алгоритъма.* За целта би било удобно да се променят дефиници 4.3.1 и 4.3.2, като се махне думата „partial" в първата, в двете се замени $\mathbb{N}$ с $\mathbb{N} \cup \{\infty\}$, а от двете равенства във втората се махне първото и се замени второто със следното

$$L(U) = \min\{c(\omega) \mid \omega \in \mathfrak{A},\ l(\omega) = U \in Inf(\mathcal{L})\},$$

като се приеме, че $\min \varnothing = \infty$ (нещо, което вече е използвано негласно в дефиниция 1.3.5). Нека $L(\alpha)$, $L_r(\alpha, j)$ и $L'_r(\alpha, j)$ са дефинирани както в глава 5, но при така изменената дефиниция за представяне на множество от подравнявания, а $L_l(\alpha, j)$ и $L'_l(\alpha, j)$ да са дефинирани по начин симетричен на онзи от дефиницията на $L_r(\alpha, j)$ и $L'_r(\alpha, j)$. В такъв случай ще бъде в сила следното:

Нека α е връх на $\mathcal{T}(V)$, а U е дума от Σ^*. Ако α е лист на $\mathcal{T}(V)$, то

$$L(\alpha)(U) = \begin{cases} 0, & \text{ако } U = V_\alpha \in Inf(\mathcal{L}), \\ \infty & \text{в противен случай.} \end{cases}$$

Нека α не е лист на $\mathcal{T}(V)$. Тогава

$$L(\alpha)(U) = \min(L_r(\alpha, N_{\alpha 1})(U), L_l(\alpha, N_{\alpha 0})(U)),$$

като при $j = 0, 1, 2, \ldots, N_{\alpha 1}$ имаме равенствата

$$\begin{aligned} L_r(\alpha, j)(U) = \min\big(\{L'_r(\alpha, j)(U)\} \cup \{&L_r(\alpha, j)(W) + c(op) \mid W \in \Sigma^*,\ op \in \Lambda, \\ &U = W \circ l(op) \in Inf(\mathcal{L}),\ L_r(\alpha, j)(W) + c(op) \le b_\alpha\}\big) \end{aligned}$$

а при $j = 0, 1, 2, \ldots, N_{\alpha 0}$ – равенствата

$$L_l(\alpha, j)(U) = \min\left(\{L'_l(\alpha, j)(U)\} \cup \{c(op) + L_l(\alpha, j)(W) \mid op \in \Lambda,\ W \in \Sigma^*,\right.$$
$$\left. U = l(op) \circ W \in Inf(\mathcal{L}),\ c(op) + L_l(\alpha, j)(W) \le b_\alpha\}\right).$$

Освен това $L'_r(\alpha, 0) = L(\alpha 0)$, $L'_l(\alpha, 0) = L(\alpha 1)$,

$$L'_r(\alpha, j)(U) = \min\{L_r(\alpha, j-1)(W) + c(op) \mid W \in \Sigma^*,\ op \in Op,$$
$$U = W \circ l(op) \in Inf(\mathcal{L}),\ r(op) = V_{\alpha 1}[j],\ L_r(\alpha, j-1)(U_0) + c(op) \le b_\alpha\},$$

при $j = 1, 2, \ldots, N_{\alpha 1}$ и

$$L'_l(\alpha, j)(U) = \min\{c(op) + L_l(\alpha, j-1)(W) \mid op \in Op,\ W \in \Sigma^*,$$
$$U = l(op) \circ W \in Inf(\mathcal{L}),\ r(op) = V_{\alpha 0}[N_{\alpha 0} + 1 - j],\ c(op) + L_l(\alpha, j-1)(U_0) \le b_\alpha\},$$

при $j = 1, 2, \ldots, N_{\alpha 0}$.

2. *Замяна на някои означения с такива, които повече да улесняват читателя.* Вместо означенията $L(\alpha)$, $L_r(\alpha, j)$, $L'_r(\alpha, j)$, $L_l(\alpha, j)$, $L'_l(\alpha, j)$, $\overrightarrow{\mathfrak{A}}^j$, $\tilde{\mathfrak{A}}^j$ и $\overleftarrow{\mathfrak{A}}^j$ по-подходящи биха били да речем съответно означенията L_α, $\overrightarrow{L}_{\alpha,j}$, $\overrightarrow{L}'_{\alpha,j}$, $\overleftarrow{L}_{\alpha,j}$, $\overleftarrow{L}'_{\alpha,j}$, $\overrightarrow{\mathfrak{A}}_{\alpha,j}$, $\overrightarrow{\mathfrak{A}}'_{\alpha,j}$ и $\overleftarrow{\mathfrak{A}}_{\alpha,j}$ (например в полза на замяната на $L(\alpha)$ с L_α говори съображението, че за всеки връх α на $\mathcal{T}$ $L(\alpha)$ е функция, чийто аргумент приема стойности от Σ^*, и по-удобно би било стойността на тази функция за дадена дума U да означаваме с $L_\alpha(U)$ отколкото с $L(\alpha)(U)$).

Качества на автореферата

Смятам, че авторефератът коректно отразява съдържанието на дисертацията. Отношение към дадената в него оценка на приносите взех в дела „Приноси в дисертацията“ на настоящата рецензия.

Заключение

Представеният дисертационен труд съдържа сериозни резултати по разглежданата в него актуална и трудна тема, които са оригинални научни приноси. Препоръчвам да бъде присъдена на автора на труда асист. Стефан Владимиров Герджиков образователната и научна степен „доктор“.

София, 3.01.2014 г. Подпис на рецензента:

Приложение към рецензията: Други критични бележки

Бележките по-долу за дисертацията на асист. Стефан Герджиков не са подредени по важност или според вида им, а общо взето са в същата последователност както местата от дисертацията, за които се отнасят.

В дефиниция 1.2.1 и на много други места са използвани означения от типа $f : A \to B$ за функции f, чиято дефиниционна област може да е същинско подмножество на A. В споменатата лема функцията, за която става дума, е наречена „partial function", но на други места липсва такова уточнение. За да няма недоразумения, щеше да е добре да се направи обща уговорка по въпроса или пък да се използват подходящи други означения. Освен това в доста от случаите щеше да е добре да се каже нещо за дефиниционната област на разглежданата функция.

Лема 1.2.5 не е вярна без допълнителни предположения. Контрапримери са да речем автоматите $(\Sigma, Q, I, \Delta, T)$, за които Q не е празно, но някое от множествата I и T е празно.

На стр. 5, ред 11 отг., пише „we appear the usual Levenshtein edit-distance", но глаголът appear не допуска пряко допълнение.

На стр. 7, ред 7 отг., вместо „$d = (0p, c)$" би трябвало да бъде „d is induced by $(0p, c)$".

В импликацията, с която завършва лема 1.5.2, липсва един знак за равенство и според мене вместо arg min трябва да бъде само min.

Първите изречения на двата абзаца след лема 1.5.2 ми се виждат неясни.

Не е изяснено достатъчно каква е връзката на понятията, въведени в параграф 1.8, и на доказаните там твърдения с понятия и резултати от цитираната в началото на параграфа книга [17]. Добре би било да се посочи например на кое понятие от книгата отговарят матриците, въведени в дефиниция 1.8.3, както и да се коментира обстоятелството, че в книгата се разглеждат формални степенни редове само с една променлива, докато в дисертацията редовете могат да бъдат с произволен краен брой променливи. Такива разяснения биха улеснили преценката за приносите в дисертацията.

Редът, въведен в дефиниция 1.8.8, може да има повтарящи се членове, така че разглеждането му като степенен ред е в донякъде разширен смисъл.

Не е изяснено в какъв смисъл дефиниция на $g_A(x)$ чрез равенството в изречението след дефиниция 1.8.8 се разглежда като еквивалентна на дефиниция 1.8.8 (аналогично е положението с дефиниция на $g_{A,n_1}(x)$ чрез равенството в изречението след дефиниция 1.8.9). Дефиниция 1.8.8 и равенството в изречението след нея дават различни формални изрази за $g_A(x)$, а в случай, че се ограничим само с въпроса за сходимостта и сумата, възниква например следният проблем: ако векторът x има компоненти с различни знаци, то редът от дефиниция 1.8.8 може да се окаже разходящ при всяка подредба на членовете му, а редът от споменатото равенство да бъде сходящ (като прост пример за това можем да посочим случая, когато $|\Sigma| = 2$, $|Q| = 1$, $|\Delta| = 2$, $x_1 = 1$, $x_2 = -1$).

Във формулировката на дефиниция 1.8.9 предположението, че $n_1 \in \mathbb{Q}$, е поставено на неподходящо място. На стр. 17, ред 13 отд., стр. 18, ред 3 отг., и по-нататък е използвано невъведено преди това означение S. На стр. 18, ред 13 отг., пише „$X \circ V \in Inf(s)$ is an infix of S“ и значи е налице излишно повторение.

Твърдението на стр. 18, че релацията $\equiv_W$, въведена в [11], съвпада с релацията $\sim_{Suf(W)}$, се нуждае от уточнение, защото първата от тези релации е дефинирана само за непразни думи и при непразна дума W. Би трябвало да се разшири обхватът на съответната дефиниция от [11] (например като се приеме $end\text{-}pos_W(X)$ да е множеството от дължините на онези начала на W, които имат край X).

На стр. 19, ред 2 отг., думата „at“ изглежда е излишна. На ред 16 отг. на същата страница вместо „a equivalence“ трябва да бъде „an equivalence“. В лема 2.1.3 вместо X трябва да бъде X_0.

Във фиг. 2.1 стрелката от състоянието [ab] към състоянието [abab] всъщност би трябвало да бъде към състоянието [aba].

На втория ред от първия абзац, започващ на стр. 20, вместо $Suf(S)$ трябва да бъде $Suf(W)$. В същия абзац с $\mathcal{A}_W$ първоначално е означен минимален автомат за езика $Suf(W)$, но след това се допуска възможност множеството на заключителните състояния на $\mathcal{A}_W$ да се избере другояче. Явно нещата във връзка с този автомат би трябвало да се изкажат по-ясно.

На стр. 21, ред 8 отд., би трябвало да има глагол след думата „we“. Думата „only“ прави невярно твърдението на същия ред, което е между думата „that“ и запетаята (контрапример е празната дума).

На стр. 34, ред 6 отд., вместо „Lemma 6.1.1“ трябва да бъде „Lemma 4.1.1“, но това, че ω_0 и ω_1 са точно посочените на следващия ред, се вижда направо, а не с помощта на лемата.

Второто изречение на лема 4.2.7 е синтактично и семантично неправилно. Подобни забележки могат да се направят и за изречения, фигуриращи в леми 4.2.9, 4.4.7, 6.1.6, 6.1.8 и в следствие 6.1.2.

В дефиниция 4.3.2 не само евентуалната неразрешимост на езика $Inf(\mathcal{L})$ може да бъде източник на неефективност. Например, ако множеството $\mathfrak{A}$ е произволно, неразрешимост на това множество също би могла да бъде причина за неефективност.

В следствие 4.3.5 би трябвало да стои знак за равенство вместо знака $\subseteq$ и $c(op)$ вместо $c(X)$. На стр. 42, ред 19 отд., вместо „$\mathcal{L}$-represents“ трябва да бъде „$\mathcal{L}$-represent“. Думата „an“ на стр. 48, ред 1 отд., е излишна. Нещо (вероятно думата „be“) липсва във второто изречение на стр. 49. На ред 9 отг. на същата страница вместо номер на глава стоят въпросителни. Пак на тази страница, на ред 12 отг., липсва думата „in“.

За по-лесното ориентиране във формулировката на лема 5.2.1 щеше да е добре да беше отбелязано изрично, че ще се отъждествяват изображението $L(\alpha)$ и графиката му (подобна бележка, но във връзка с по-общ вид изображения е направена едва по-късно – в бележката под линия на стр. 54). Вместо „$\{\langle U, 0\rangle\}$ *if* $U = V_\alpha$ *and* $V_\alpha \in Inf(\mathcal{L})$“ би трябвало да се напише просто „$\{\langle V_\alpha, 0\rangle\}$ *if* $V_\alpha \in Inf(\mathcal{L})$“. В лемата фигурира означението N без да е посочен смисълът му.

На стр. 53, ред 19 отд., вместо „form“ трябва да бъде „from“.

Описанието на алгоритъма за търсене, дадено в параграф 5.3.1, е непълно и е трудно за следене. В частност не е обяснено какъв е смисълът на $cost(u)$ и $act(u)$, а също как се представя $L_r(\alpha, 0)$. Във връзка с последното смятам, че би било подходящо първо да се дефинира как се представя произволно изображение на крайно подмножество на $Inf(\mathcal{L})$ в $\mathbb{N}$ и по този начин да стане ясно как се представят всички разглеждани $L_r(\alpha, j)$. В словесното описание на алгоритъма никъде не се среща случаят, когато за даден връх u трябва да се положи $act(u) = false$. Според мене е излишно обръщания на думи и операции да се включват като стъпки на този алгоритъм при разширение наляво – биха могли да се използват действия, симетрични на действията при разширение надясно, а обръщанията да се използват евентуално само в някои разсъждения относно алгоритъма.

Изглежда написаното на стр. 53, редове 11 и 10 отд., „the subquery (abb, 1) in the regular language $\{\mathsf{ababb}, \mathsf{acbbb}\}^*$" означава, че се търси множеството $\{U \in Inf(\{\mathsf{ababb}, \mathsf{acbbb}\}^*) \,|\, d(U, \mathsf{abb}) \leq 1\}$, но тук би бил по-подходящ пример с неравенството $d(U, \mathsf{abb}) \leq q|\mathsf{abb}|$ при някое избрано рационално число q между 0 и 1.

На фиг. 5.2 и 5.3 някои прави букви са заменени с курсивни. Езикът на автомата, изобразен на фигури 5.2 и 5.3, се различава от посочения там език $Inf(\{\mathsf{ababb}, \mathsf{acbbb}\}^*)$. Например думите cbbbab и cbbbac от езика $Inf(\{\mathsf{ababb}, \mathsf{acbbb}\}^*)$ не принадлежат на езика на въпросния автомат, а думите bc и bbc от езика на автомата, не принадлежат на $Inf(\{\mathsf{ababb}, \mathsf{acbbb}\}^*)$.

В написаното под фиг. 5.2 едното от крайните „s" в „lists" и „represents" представлява грешка от граматическа гледна точка, а означението $L_r(\mathsf{abb}, 1)$ няма смисъл (първият аргумент на L_r би трябвало да бъде дума от нули и единици, а освен това би трябвало предварително да са избрани дума V и рационално число q между 0 и 1, за които да се търси множеството $\{U \in Inf(\{\mathsf{ababb}, \mathsf{acbbb}\}^*) \,|\, d(U, V) \leq q|V|\}$.)

На стр. 55 и по-нататък би било по-добре вместо $c(u)$ да стои c. На същата страница, ред 17 отг., вместо „alignment" би трябвало да стои „alignments". По-добре би било дясната страна на дефиниционното равенство на множеството $S(j-1, l)$ да се напише във вида $\{n_1 + l, n_2 + l, \ldots, n_{M_{j-1}} + l\}$. На ред 14 отд. вместо „the sets $S(j-1, op)$" трябва да бъде „the union of the sets $S(j-1, l)$". На ред 10 отд. вместо „every n and every operation $l \leq \mu$" би трябвало да бъде „every k and every $l \leq \mu$".

В лема 5.3.1 и нейното доказателство вертикалните черти пред и след M_{j-1} са излишни.

На стр. 59, ред 11 отг., има излишно „one". На ред 18 отг. е пропусната една затваряща квадратна скоба. На ред 18 отд. неточно се споменава за вмъкване на u в $B_\alpha^j[n]$ – всъщност би трябвало да става дума за вмъкване на някоя двойка с пръв член u.

Първото равенство в доказателството на лема 5.3.5 може да се напише по-просто така: $label(u') = label(u)$ if $(u, c(u)) \in L_r(\alpha, N_{\alpha 1}) \cup L_l(\alpha, N_{\alpha 0})$.

На стр. 68, ред 18 отд., „an" трябва да се махне или да се замени с „two".

Позоваването на лема 6.1.3 в доказателството на лема 5.3.6 е явно резултат на някаква грешка.

На стр. 71, редове 19-20 отг., фигурира странното словосъчетание „the approximate search problem in for languages".

Не забелязах да е обяснено какво означава $W_{\mathcal{L}}$ в доказателството на лема 5.3.8.

На стр. 85, ред 17 отд., вместо $\rho(Op) - 1$ трябва да стои $\rho(Op)$, а тогава онова, което се твърди в изречението, започващо на този ред, няма да е вярно.

В следствие 6.1.2 вместо $c(\omega_0)$ и $c(\omega_1)$ би трябвало да стои съответно $c(\omega_0')$ и $c(\omega_1')$. В доказателството на следствието трябва да се добави „the right side of" пред „the last operation".

В равенството, което е в заключението на лема 6.1.5, вместо $k = 1$ трябва да стои $k = 0$, защото в противен случай твърдението на лемата се опровергава с контрапример. Нека $\Sigma = \{a\}$, а (Op, c) е Левенщайновото ортографско разстояние и значи имаме $\rho = 1$. Нека $V_0 = V_1 = a$, $V = aa$, $b_0 = b_1 = 1/2$, $b = 1$. Тогава $(\varepsilon, a)(a, a) \in \mathfrak{A}^{\leq b}(V)$. От друга страна обаче

$$\bigcup_{k=0}^{\rho-1} \overrightarrow{\mathfrak{B}}_k \cup \bigcup_{k=1}^{\rho-1} \overleftarrow{\mathfrak{B}}_k = \overrightarrow{\mathfrak{B}}_0,$$

а в случая всяко подравняване от множеството $\overrightarrow{\mathfrak{B}}_0$, има начало, принадлежащо на $\mathfrak{A}^{\leq b_0}(a)$, т.е. има пръв член $\langle a, a \rangle$.

За да има смисъл дефиницията на множествата $\overrightarrow{\mathfrak{B}}_k$ и $\overleftarrow{\mathfrak{B}}_k$, дадена в същата лема, би трябвало k да е естествено число и да имаме съответно неравенствата $k \leq n_0$ и $k \leq n_1$, но такова ограничение не е формулирано, когато се дава тази дефиниция (същите ограничения са нужни и за двете равенства в края на доказателството на лемата).

При дефинирането на множествата $\overrightarrow{\mathfrak{B}}_k^j$ в лема 6.1.6 са нужни още предположения за k и j освен неравенството $k + j \geq 0$ – например това, че k е естествено число, удовлетворяващо неравенството $k \leq n_0$, а j е цяло число и $j \leq n_1$. В заключението на лемата фигурира означението $\overrightarrow{\mathfrak{B}}_k^{j'}$ при j', менящо се от $j - \rho$ до $j - 1$. При $\rho > 1$, $k < \rho$ и $1 \leq j < \rho - k$ обаче са в сила неравенствата $j - \rho \leq -k - 1 < j - 1$, а споменатото означение няма смисъл за стойностите на j' от $j - \rho$ до $-k - 1$ (за тях $k + j' < 0$). Вероятно трябва да се разшири дефиницията на въпросните множества, като се приеме, че те са празни при $k + j < 0$. Аналогични бележки могат да се направят и за дефиницията на множествата $\overleftarrow{\mathfrak{B}}_k^j$ в лема 6.1.8.

На стр. 88, ред 2 отд., липсва думата „for", но всъщност написаното на този ред би могло просто да се редуцира до първата дума на реда.

В доказателството на лема 6.1.6 на два пъти се среща изразът „for some $j - \rho \leq j' < j$", а би трябвало да бъде „for some integer j' satisfying the inequalities $j - \rho \leq j' < j$". Разглеждането в това доказателство на случая, когато $\omega_1 = \varepsilon$, е излишно усложнено – в този случай всъщност имаме $j' = -k$.

Това, което се твърди в последното изречение от доказателството на лема 6.1.8 е неосновователно – при $k = 0$ свеждането към лема 6.1.6 няма да върви поради това, че дефиницията на множеството $\overrightarrow{\mathfrak{B}}_k^j$ при $k = 0$ е различна от дефиницията на това множество при $k > 0$, а няма такова различие при дефиницията на множеството $\overleftarrow{\mathfrak{B}}_k^j$.

Изречението на стр. 92, ред 11 отг., е неясно (струва ми се, че авторът е искал да каже нещо друго).

Не е ясно защо функциите, въведени чрез дефиниция 7.1.1, са наречени „generating functions". Терминът „generating function" обикновено се употребява като синоним на „formal power series", а изразите за конкретните функции, за които става дума, са суми от краен брой събираеми и имат стойност за всички стойности на променливите с изключение евентуално на нулеви стойности на $t_1, \ldots, t_\Sigma$ (в случая на функциите f_i и f_Σ).

На стр. 115, редове 14 и 13 отд., „is actually of the form" не са подходящите думи за случая.

За да има смисъл формулировката на свойството 1 в лема 7.1.4, би трябвало в дефиниция 7.1.3 вместо $\mathbb{N}^{|\Sigma|}$ да стои $\mathbb{Z}^{|\Sigma|}$ (в такъв случай, разбира се, бихме имали $\mathcal{A}(U,q;s) = \varnothing$ и $a(U,q;s) = 0$ при $s \in \mathbb{Z}^{|\Sigma|} \backslash \mathbb{N}^{|\Sigma|}$). В споменатата дефиниция би било по-добре да се използва A вместо $\mathcal{A}$, защото въведеното означение $\mathcal{A}(U,q)$ се среща по-нататък в доказателството на лема 7.1.6, а там $\mathcal{A}$ е краен автомат.

В общия случай може да се окаже, че равенството в свойството 2 от лема 7.1.4 съдържа изрази, нямащи смисъл за някои вектори t с нулеви компоненти. Поради това би било добре във формулировката на въпросното свойство да се поиска всички компоненти на разглежданите вектори t да бъдат различни от 0. За да не възниква въпросът дали сходимостта и сумата на реда в лявата страна на равенството не зависят от подредбата на членовете, още по-добре е да поискаме всички тези компоненти да бъдат положителни (освен това е лесно да се види, че без току-що изказаното изискване неравенството $f_\varepsilon(z^{-q}t; z) < 1$, фигуриращо в свойството 2, не е достатъчно за сходимостта на въпросния ред). Понеже това ограничение фигурира в статията [19] на автора, предполагам, че то просто е пропуснато случайно тук.

В доказателствата на леми 7.1.4 и 7.1.6 (особено в първото) се използва широко техниката на смятане с формални степенни редове, но не е цитирана никаква литература, в която се излага тази техника. При това формални степенни редове се преобразуват в изрази, които не са формални степенни редове в буквалния смисъл на думата, а пък се третират като такива (имам предвид например изразите, които са вдясно от знаците за равенство на стр. 119, редове 12-13 отг. и редове 7-6 отд., както и израза на стр. 120, ред 3 отг.). В доказателството на лема 7.1.6 би могло и да не се прибягва до тази техника, защото тук става дума само за сумите (евентуално безкрайни) на редове с неотрицателни членове, но няма как да се проследи детайлно доказателството на лема 7.1.4 без читателят да използва достатъчно много допълнителни сведения във връзка с въпросната техника (вероятно и свойствата на въведената в дефиниция 7.1.2 частична наредба би трябвало да бъдат разгледани по-подробно преди да се пристъпи към доказателството).

На стр. 118, ред 17 отд., изречението в началото на реда е неточно. В частта от доказателството, която е на тази страница, знаците „прим“ и „секонд“ би трябвало да бъдат не към символа op, а към индексите след него. На стр. 119, ред 2 отд., „and $l(\omega_\varepsilon) = \varepsilon$“ е излишно. Не е ясно каква роля играят неравенството в квадратните скоби и думата „and“ на стр. 120, ред 8 отг. Горната граница k при знак за сумиране и знаци за умножаване трябва да се замени с m навсякъде на страници 121 и 122 с изключение на последния път, където тя трябва да се замени с n. Всяко от условията от вида $l(op_j) \in Op_i$ на стр. 122 трябва да стане $op_j \in Op_i$. На редове 13, 11 и 4 отд. вместо j_i трябва да бъде i_j. Вместо конюнкцията на ред 5 отг. и повторенията ѝ в сумационни индекси след това би било по-добре да се напише само вторият ѝ член (поправен в съответствие с горната забележка), но с указание, че той трябва да е в сила при $j = 1, \ldots, m$. Добре ще е да се добави също $j = 1, \ldots, n$ в сумационния индекс на ред 4 отд.

На стр. 124, редове 7 и 10 отг., вместо „properties“ и „diverge“ трябва да бъде съответно „property“ и „diverges“. От двете последователни „the“ – в края на ред 10 отг. и в началото на следващия ред – едното е излишно. Излишно е и „there“ в края на ред 14 отг. На ред 16 отг., ред 9 отд. и ред 5 отд. вместо $\mathcal{L}$, $gen_{\mathcal{L}(\mathcal{A})}(V)$ и $a(U; s)$ трябва да бъде съответно $\mathcal{L}(\mathcal{A})$, $gen_{\mathcal{L}(\mathcal{A})}(V, q)$ и $a(U, q; s)$.

В лема 7.1.6 би било по-добре да се положи например $g_{\mathcal{A},i} = \frac{\partial g_\mathcal{A}}{\partial t_i}$ при $i = 1, \ldots, |\Sigma|$ и да се напише $g_{\mathcal{A},i}(v(t; z))$ вместо $\frac{\partial g_\mathcal{A}}{\partial v_i}(v(t; z))$ във второто неравенство в заключението на лемата. Доказателството на неравенството би станало по-прегледно, ако предварително беше доказано, че при $t \in \mathbb{R}_+^{|\Sigma|}$ е в сила неравенството

$$\sum_{U \in Inf(\mathcal{L}(\mathcal{A}))} |U| t^{\|U\|} \leq \sum_{i=1}^{|\Sigma|} t_i \frac{\partial g_\mathcal{A}}{\partial t_i}.$$

Вместо $v^{\|U\|}(t; z)$ би било по-добре навсякъде в доказателството да се пише $v(t; z)^{\|U\|}$.

Второто от неравенствата на стр. 125, ред 2 отд., предполага, че $\frac{1}{\rho} \geq q$, без това да е отбелязано в разсъжденията, които се правят. В следствие 7.1.7 би било добре да се добави индекс n_1 също в означенията $gen_{\mathcal{L}(\mathcal{A})}(V, q)$ и $Gen_{\mathcal{L}(\mathcal{A})}(V, q)$.

Не е ясно какво се има предвид във формулировката и доказателството на лема 7.2.1, когато се използва означението $pr(V | V \in \Sigma^N)$. Изглежда, че при $V \in \Sigma^N$ то означава просто $pr(V)$, но в такъв случай каква е нуждата от това означение, ако се разглеждат само такива V, които са от Σ^N?

В равенството на стр. 133, ред 2 отг., вместо n трябва да бъде $|Q|$. В изречението, съдържащо това равенство, става дума за матрица $M^\bullet_\mathcal{A}(t)$, която според написаното била дефинирана в параграф 1.8. Това не е съвсем точно – поне в явен вид такава матрица не е разглеждана във въпросния параграф. Не е съвсем точно и твърдението, че там е установено споменатото равенство – по същество тук е направена допълнително и една промяна в реда на сумирането. Остава и отбелязаният по-рано в това приложение въпрос за смисъла на подобни равенства.

В доказателството на лема 7.3.3 не е посочено, че $n = |Q|$. В първото изречение от доказателството вместо номер на параграф стоят въпросителни.

В следствие 7.3.4 вместо $\delta(p, \sigma_i)$ трябва да бъде $\delta(k, \sigma_i)$.

На стр. 137, ред 8 отд., вместо „the for" трябва да бъде само „for".

Параграф 7.3 оставя впечатление за недовършеност. Целта, която се преследва, е да се покаже, че в определен достатъчно естествен клас от случаи средното време за работа на алгоритъма от предходните глави има оценка отгоре, която е линейна относно дължината на думата, към която се прилага алгоритъмът. В предходните параграфи и в този параграф е извършена сериозна подготвителна работа, но прилагането на резултатите от нея за постигането на въпросната цел е описано само разказвателно в последните 13 реда от параграфа без да е формулирана теорема, съдържаща явно описание на клас, за какъвто става дума. Опитите на читателя сам да попълни тази празнота са затруднени например от обстоятелството, че оценката в общия случай при $\rho(Op) = 1$, съдържаща се в твърдение 7.2.7, дава линейна оценка тогава, когато са сходящи посочените в това твърдение редове $g_{\mathcal{A}}(v^{(2)}(t;z))$ и $\frac{\partial g_{\mathcal{A}}}{\partial v_i}(v(t;z))$, а следствие 7.3.5, чрез което би трябвало да се получи желаната теорема, е изказано като достатъчно условие да бъде сходящ ред $g_{\mathcal{A}}(v(t;z))$. Все пак трудността изглежда преодолима, защото редът $\frac{\partial g_{\mathcal{A}}}{\partial v_i}(v(t;z))$ вероятно случайно е пропуснат в следствие 7.3.5, а функцията $v^{(2)}$ е от вида на функцията v, но с двойно по-голяма стойност на параметъра q.

В дефиниция 8.1.5 вместо „with root V" би трябвало да бъде „with root labelled by V", защото в противен случай би могло някой път в условията на точка 2 да се наложи дърветата $\mathcal{T}_{\mathcal{W}}^{(l)}$ и $\mathcal{T}_{\mathcal{W}}^{(r)}$ да имат общ връх. И в дефиниция 8.2.1 вместо „with root V" би трябвало да бъде „with root labelled by V".

На стр. 143, ред 15 отд., вместо „be decompose" трябва да бъде „be decomposed".

На стр. 173 и в библиографския списък е посочен номер 1319 на последната страница на статията на дисертанта в „Доклади на БАН" вместо верния номер 1318. В библиографския списък не е ясен принципът на подреждане на онези източници, за които информацията е само Интернет линк – част от тях са с номера от 1 до 6 без да са азбучно подредени, а един е с номер 62. В pdf файла на дисертацията тези линкове са активни, но използването на един от тях води до опит за отваряне на несъществуваща Интернет страница на адрес http://trec.nist.gov/pubs/trec5/t5, вместо да отваря съответния източник.

София, 3.01.2014 г. Подпис на рецензента: