Aspects algorithmiques et combinatoires de la théorie des automates

Partie II: analyse en moyenne de l'algorithme de Glushkov

Cyril Nicaud, LIGM, Paris-Est

15 juillet 2016

Dans cette partie on va reprendre la construction de l'automate de Glushkov, qui nous a servi à démontrer un des sens du théorème de Kleene. Cette fois on va se placer plus dans les thématiques de l'école et chercher à faire une analyse en moyenne de la complexité en espace de cette construction. On se pose donc la question suivante : "Si on prend une expression rationnelle aléatoire, combien l'automate de Glushkov a-t-il de transitions en moyenne ?"

Dans toute la suite on considère uniquement l'alphabets de taille 2 $A = \{a, b\}$, pour simplifier les écritures. Cependant, tout s'étend directement à n'importe quelle taille d'alphabet.

1 Expressions rationnelles aléatoires

Pour notre modèle, on voit une expression comme un arbre unaire-binaire, dont les nœuds binaires sont étiquetés par $+$ ou $\bullet$, les nœuds unaires sont étiquetés par $\star$ et les feuilles par a, b ou ε. La taille d'un arbre est son nombre de nœuds (internes + feuilles).

1.1 Série génératrice

On a donc la spécification suivante pour l'ensemble $\mathcal{R}$ des arbres d'expressions rationnelles :

$$\mathcal{R} = \circ_\varepsilon + \circ_a + \circ_b + \overset{\star}{|}_{\mathcal{R}} + \overset{+}{\wedge}_{\mathcal{R}\ \mathcal{R}} + \overset{\bullet}{\wedge}_{\mathcal{R}\ \mathcal{R}}. \tag{1}$$

On en déduit que la série génératrice associée $R(z)$ satisfait

$$R(z) = 3z + zR(z) + 2zR(z)^2.$$

On résoud cette équation, on identifie la bonne solution, et on trouve :

$$R(z) = \frac{1 - z - \sqrt{\Delta(z)}}{4z}, \text{ avec } \Delta(z) = (1 - z/\rho)(1 - z/\tilde{\rho}) \text{ et } \begin{cases} \rho = \frac{1}{1+2\sqrt{6}} \approx 0.17 \\ \tilde{\rho} = \frac{1}{1-2\sqrt{6}} \approx -0.26 \end{cases} \tag{2}$$

En appliquant le théorème de transfert, on trouve donc l'asymptotique du nombre d'expressions :

$$R(z) \sim_{z=\rho} \frac{1-\rho}{4\rho} - \frac{\sqrt{1-\rho/\tilde{\rho}}}{4\rho}\sqrt{1-z/\rho} \ \Rightarrow\ [z^n]R(z) \sim A \cdot \frac{\rho^{-n}}{n^{3/2}},$$

avec

$$A = \frac{\sqrt{1-\rho/\tilde{\rho}}}{8\rho\sqrt{\pi}} \approx 0.54$$

1.2 Nombre moyen de feuilles

On sait que le nombre d'états de l'automate de Glushkov est le nombre de lettres de l'expression plus 1. Dans cette partie, on essaye donc de quantifier l'espérance du nombre de lettres pour avoir une première indication sur la taille de l'automate.

Pour cela on va utiliser la technique du marquage, et des séries bivariées, en ajoutant une marque sur les feuilles étiquetées par une lettre. La spécification devient :

$$\overline{\mathcal{R}} = \circ_\varepsilon + \overline{\circ_a} + \overline{\circ_b} + \overset{\star}{|}_{\overline{\mathcal{R}}} + \overset{+}{/\!\backslash}_{\overline{\mathcal{R}}\,\overline{\mathcal{R}}} + \overset{\bullet}{/\!\backslash}_{\overline{\mathcal{R}}\,\overline{\mathcal{R}}}. \tag{3}$$

On introduit une autre variable formelle u, de sorte que le coefficient en $z^n u^k$ soit le nombre d'expressions de taille n avec k lettres. La méthode symbolique s'applique encore, il faut juste multiplier par u à chaque fois qu'on crée un nœud qui est une lettre. Cela donne :

$$R(z,u) = z + 2uz + zR(z,u) + 2zR(z,u)^2.$$

On pourrait suivre le processus classique : résoudre, dériver par rapport à u, évaluer en $u = 1$ pour calculer l'espérance du nombre de lettres. Mais ici, c'est un peu plus simple de d'abord dériver et évaluer en $u = 1$ avant de résoudre :

$$\frac{d}{du}R(z,u) = 2z + \frac{d}{du}R(z,u) + 4zR(z,u)\cdot\frac{d}{du}R(z,u),$$

et donc, en utilisant le fait que $R(z,1) = R(z)$, on a

$$\frac{d}{du}R(z,u)\Big|_{u=1} = \frac{2z}{1-z-4zR(z)} = \frac{2z}{\sqrt{\Delta(z)}},$$

la dernière égalité découlant directement de l'équation (2).

On applique le théorème de Transfert près de singularité dominante :

$$\frac{d}{du}R(z,u)\Big|_{u=1} \sim_{z=\rho} \frac{2\rho}{\sqrt{1-\rho/\tilde{\rho}}}\cdot\frac{1}{\sqrt{1-z/\rho}} \;\Rightarrow\; [z^n]\frac{d}{du}R(z,u)\Big|_{u=1} \sim B\cdot\frac{\rho^{-n}}{\sqrt{n}}$$

avec

$$B = \frac{2\rho}{\sqrt{\pi(1-\rho/\tilde{\rho})}}.$$

On en déduit l'asymptotique de l'espérance du nombre de lettres :

$$\mathbb{E}[\sharp\text{lettres}] = \frac{[z^n]\frac{d}{du}R(z,u)\Big|_{u=1}}{[z^n]R(z)} \sim \frac{B}{A}\cdot n = \frac{24-2\sqrt{6}}{69}\,n \approx 0.28\,n.$$

En particulier il y a un nombre moyen linéaire de lettres dans une expression aléatoire uniforme.

1.3 Reconnaissance du mot vide

Nous allons maintenant calculer la probabilité asymptotique qu'une expression rationnelle aléatoire reconnaisse le mot vide. On note $\mathcal{R}_\varepsilon$ l'ensemble des arbres dont l'expression reconnaît le mot vide et $\mathcal{R}_{\overline{\varepsilon}}$ ceux qui ne reconnaissent pas le mot vide. On a la spécification suivante :

$$\mathcal{R}_\varepsilon = \circ_\varepsilon + \overset{\star}{|}_{\mathcal{R}} + \overset{+}{/\!\backslash}_{\mathcal{R}_{\overline{\varepsilon}}\,\mathcal{R}_\varepsilon} + \overset{+}{/\!\backslash}_{\mathcal{R}_\varepsilon\,\mathcal{R}_{\overline{\varepsilon}}} + \overset{+}{/\!\backslash}_{\mathcal{R}_\varepsilon\,\mathcal{R}_\varepsilon} + + \overset{\bullet}{/\!\backslash}_{\mathcal{R}_\varepsilon\,\mathcal{R}_\varepsilon}$$

On applique la méthode symbolique et on obtient, pour les séries génératrices $R_\varepsilon(z)$ et $R_{\overline{\varepsilon}}(z)$:

$$R_\varepsilon(z) = z + zR(z) + 2zR_{\overline{\varepsilon}}(z)R_\varepsilon(z) + 2zR_\varepsilon(z)^2.$$

En utilisant le fait que $R_\varepsilon(z) + R_{\overline{\varepsilon}}(z) = R(z)$, on obtient :

$$R_\varepsilon(z) = z + zR(z) + 2z\big(R(z) - R_\varepsilon(z)\big)R_\varepsilon(z) + 2zR_\varepsilon(z)^2 = z + zR(z) + 2zR(z)R_\varepsilon(z)$$

Par conséquent,

$$R_\varepsilon(z) = \frac{z + zR(z)}{1 - 2zR(z)} = \frac{(1 + 3z - \sqrt{\Delta(z)})(1 + z - \sqrt{\Delta(z)})}{8z(1 + 6z)} \tag{4}$$

Quand on développe le numérateur on trouve qu'il existe une fonction $E(z)$ qui est analytique dans un disque de rayon strictement supérieur à ρ telle que

$$R_\varepsilon(z) = E(z) - \frac{1 + 2z}{4z(1 + 6z)} \cdot \sqrt{\Delta(z)}.$$

Au passage, l'équation (4) permet de voir qu'il n'y a pas de singularité en $-1/6$ (un argument combinatoire permet de l'affirmer également). La singularité dominante est toujours ρ et en son voisinage on a

$$R_\varepsilon(z) \sim_{z=\rho} E(\rho) - \frac{1 + 2\rho}{4\rho(1 + 6\rho)} \sqrt{1 - \rho/\tilde{\rho}} \cdot \sqrt{1 - z/\rho}.$$

On applique le théorème de transfert :

$$[z^n]R_\varepsilon(z) = \frac{1 + 2\rho}{1 + 6\rho} \cdot \frac{\sqrt{1 - \rho/\tilde{\rho}}}{8\rho\sqrt{\pi}} \cdot \frac{\rho^{-n}}{n^{3/2}}$$

Et donc on a

$$\mathbb{P}_n(\text{reconnaît le mot vide}) \xrightarrow[n\to\infty]{} \frac{1 + 2\rho}{1 + 6\rho} \approx 0.664$$

2 Algorithme de Glushkov

2.1 Taille de First

Pour commencer à étudier le nombre de transitions moyen dans l'automate de Glushkov, nous étudions le nombre de transitions qui relient l'état initial à un autre état. Autrement dit, nous sommes intéressés par le nombre de lettres (dans l'expression linéarisée) qui commencent un mot du langage de l'expression.

Pour une expression $R \in \mathcal{R}$, on va noter $\texttt{first}(R)$ le nombre de lettres qui commencent un mot de $\mathcal{R}$. On cherche à calculer la valeur moyenne de `first`, on introduit donc la série bivariée $F(z, u)$ associée au paramètre définie par

$$F(z, u) = \sum_{R \in \mathcal{R}} u^{\texttt{first}(R)}\, z^{|R|}$$

Pour la spécification, on aura également besoin des restrictions $F_\varepsilon(z)$ et $F_{\overline{\varepsilon}}(z)$ aux expressions de $\mathcal{R}_\varepsilon$ et $\mathcal{R}_{\overline{\varepsilon}}$, respectivement :

$$F_\varepsilon(z) = \sum_{R \in \mathcal{R}_\varepsilon} u^{\texttt{first}(R)}\, z^{|R|}; \quad F_{\overline{\varepsilon}}(z) = \sum_{R \in \mathcal{R}_{\overline{\varepsilon}}} u^{\texttt{first}(R)}\, z^{|R|}$$

Pour faire la spécification, il faut voir comment calculer `first` récursivement. On se convainc facilement que :

$$\begin{cases} \texttt{first}(\circ_\varepsilon) = 0 \\ \texttt{first}(\circ_a) = 1,\ \forall a \in A \\ \texttt{first}\left(\overset{\star}{|}_R\right) = \texttt{first}(R) \\ \texttt{first}\left(\overset{+}{\wedge}_{R_1\ R_2}\right) = \texttt{first}(R_1) + \texttt{first}(R_2) \\ \texttt{first}\left(\overset{\bullet}{\wedge}_{R_1\ R_2}\right) = \begin{cases} \texttt{first}(R_1) + \texttt{first}(R_2) & \text{si } \varepsilon \in R_1 \\ \texttt{first}(R_1) & \text{sinon} \end{cases} \end{cases}$$

On part donc de la spécification suivante de $\mathcal{R}$:

$$\mathcal{R} = \circ_\varepsilon + \circ_a + \circ_b + \overset{\star}{|}_{\mathcal{R}} + \overset{+}{\wedge}_{\mathcal{R}\ \mathcal{R}} + \overset{\bullet}{\wedge}_{\mathcal{R}_\varepsilon\ \mathcal{R}} + \overset{\bullet}{\wedge}_{\mathcal{R}_{\overline{\varepsilon}}\ \mathcal{R}}.$$

On va marquer les feuilles qui correspondent à des lettres qui commencent un mot de l'expression linéarisée. On obtient l'équation :

$$F(z,u) = z + 2zu + zF(z,u) + zF(z,u)^2 + zF_\varepsilon(z,u)F(z,u) + zF_{\overline{\varepsilon}}(z,u)R(z).$$

Comme $F_\varepsilon(z,u) + F_{\overline{\varepsilon}}(z,u) = F(z,u)$ on a

$$F(z,u) = z + 2zu + zF(z,u) + zF(z,u)^2 + zF_\varepsilon(z,u)F(z,u) + zF(z,u)R(z) - zF_\varepsilon(z,u)R(z).$$

On dérive par rapport à u et on évalue en $u = 1$, en notant $Q(z) = \frac{d}{du}F(z,u)\big|_{u=1}$:

$$\begin{aligned} Q(z) &= 2z + zQ(z) + 2zQ(z)R(z) + zQ_\varepsilon(z)R(z) + zR_\varepsilon(z)R(z) + zQ(z)R(z) - zQ_\varepsilon(z)R(z) \\ &= 2z + zQ(z) + 2zQ(z)R(z) + zR_\varepsilon(z)R(z) + zQ(z)R(z) \end{aligned}$$

Et donc

$$Q(z) = \frac{2z}{1 - z - 3zR(z) - zR_\varepsilon(z)}$$

Comme on a les expressions de $R(z)$ et de $R_\varepsilon(z)$, on en déduit une expression de $Q(z)$ au voisinage de la singularité de la forme :

$$Q(z) \sim_{z=\rho} H(\rho) - (12 + 4\sqrt{6}) \cdot \sqrt{\Delta(z)}$$

On applique à nouveau le théorème de transfert et on obtient la formule asymptotique de l'espérance de `first` :

$$\mathbb{E}_n[\texttt{first}] = \frac{[z^n]Q(z)}{[z^n]R(z)} \xrightarrow[n\to\infty]{} 12 + 4\sqrt{6} \approx 21.8$$

En particulier on vient de démontrer que :

Lemme 1 *L'espérance du nombre de lettre qui commencent une expression rationnelle aléatoire, une fois linéarisée, tend vers une constante quand la taille de l'expression tend vers l'infini.*

2.2 Et la suite ?

On peut analyser, en gros de la même manière, l'espérance du nombre de transitions dans l'automate de Glushkov. Le résultat est le suivant :

Théorème 2 *Le nombre moyen de transitions dans l'automate de Glushkov d'une expression rationnelle aléatoire de taille n est en $\Theta(n)$.*

Dans le pire cas, il y a $\Theta(n^2)$ transitions, le cas moyen est donc beaucoup mieux que ce qu'on pourrait anticiper d'une analyse dans le pire des cas.