Reconstruction de mots par requêtes avec poids

Gwenaël Richomme, Matthieu Rosenfeld

keywords: Combinatoire des mots; reconstruction; algorithmique du texte

Contexte: On dit qu'un mot u est un *sous-mot* d'un mot w si u et w peuvent être décomposés sous la forme $u = u_1 \cdots u_\ell$ et $w = v_0 u_1 v_1 \cdots u_\ell v_\ell$ où les $u_1, \ldots, u_\ell, v_0, \ldots, v_\ell$ sont des mots (par exemple, $foule$ est un sous-mot de $formule$). Étant donné deux mots u et v, $\binom{u}{v}$ compte le nombre d'occurrences de v dans u en tant que sous mot. Par exemple, $\binom{01101}{01} = 4$.

Une question algorithmique naturelle est de déterminer la quantité nécessaire d'information pour reconstruire un objet combinatoire. Dans le contexte des mots, de nombreuses questions restent ouvertes autour du sujet. L. O. Kalashnik [3] a introduit le problème suivant en 1973:

Question 1. *Quelle est la plus petite valeur de ℓ telle que tout mot w de longueur n peut être reconstruit à partir de toutes les valeurs $\binom{w}{u}$ pour l'ensemble des mots u de longueur ℓ ?*

La meilleure borne supérieure sur $\ell(n)$ est $\lfloor \frac{16}{7}\sqrt{|n|} \rfloor + 5$ alors que la meilleure borne inférieure est $3^{(\sqrt{2/3}-o(1))\log_3^{1/2}(|n|)}$ [4, 1]. Ce problème semble très difficile, ce qui a motivé l'étude de plusieurs variantes.

Dans la version introduite par Fleischmann *et al.*, l'objectif est de reconstruire un mot w qui n'est connu que par un oracle [2]. Dans cette tâche, on peut poser des questions à l'oracle sous une certaine forme pour réussir à déterminer le mot de manière unique. Plus précisément, à chaque tour, on choisit un mot u en fonction des réponses précédemment obtenues et on demande la valeur de $\binom{w}{u}$. L'objectif est de minimiser le nombre de questions posées. Fleischmann *et al.* ont montré qu'il suffit de $\lfloor n/2 \rfloor + 1$ questions pour reconstruire n'importe quel mot de longueur n. Nous avons montré qu'en fait $O(\sqrt{n \log n})$ questions suffisent, et qu'on peut même le faire en moyenne en $O(\log n)$ questions [5]. Nous n'avons aucune borne inférieure sur le nombre de questions à poser, donc ces résultats pourraient toujours être loin de l'optimal.

Stage: L'objectif principal du sujet proposé est l'étude de la variante suivante de ce dernier problème. Le cadre est toujours de reconstruire un mot inconnu w, en posant des questions de la forme $\binom{w}{u}$ à un oracle qui connait le mot. Mais, plutôt que de minimiser le nombre de questions, on souhaite minimiser la somme des $|u|$. L'idée étant que les questions n'ont pas toutes le même coût et en l'occurrence, plus u est long plus la question $\binom{w}{u}$ est couteuse. D'autres variantes du problème, déjà considérées dans la littérature, pourront aussi être abordées. Elles consistent, par exemple, à poser des questions sur l'existence d'occurrences plutôt que sur le nombre d'occurrences ou sur les facteurs plutôt que sur les sous-mots de w.

Ce stage se déroulera au LIRMM à Montpellier au sein de l'équipe ESCAPE. Il pourra être rémunéré si les statuts le permettent. Le travail effectué pendant ce stage pourra se poursuivre lors d'une thèse.

References

[1] M. Dudik and L.J. Schulman. Reconstruction from subsequences. *J. Combin. Theory Ser. A*, 103:337–348, 2003.

[2] P. Fleischmann, M. Lejeune, F. Manea, D. Nowotka, and M. Rigo. Reconstructing words from right-bounded-block words. *Internat. J. Found. Comput. Sci.*, 32(6):619–640, 2021.

[3] L.O. Kalashnik. The reconstruction of a word from fragments. In *Numerical Mathematics and Computer Technology, Preprint IV*, pages 56–57. Akad. Nauk. Ukrain. SSR Inst. Mat., 1973.

[4] I. Krasikov and Y. Roditty. On a reconstruction problem for sequences. *J. Combin. Theory Ser. A*, 77:344–348, 1997.

[5] G. Richomme and M. Rosenfeld. Reconstructing Words Using Queries on Subwords or Factors. In *(STACS 2023)*, volume 254 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 52:1–52:15, 2023.