Sujet de stage: recherche de motifs dans les séquences ordonnées

Julien David (Université de Caen)
Thierry Lecroq (Université de Rouen)

Printemps 2023

Ce stage est financé par la fédération de recherche NormaSTIC. Il est partagé entre l'université de Rouen ou l'université de Caen (avec un unique lieu de résidence pour la/le stagiaire). Possibilité de continuer ce sujet de recherche dans le cadre d'un doctorat.

Contact : `julien.david@unicaen.fr` et `thierry.lecroq@univ-rouen.fr`, merci de nous mettre tous les deux en copies des messages.

On s'intéresse ici à la recherche de motif dans les séries temporelles et, plus largement dans des listes de valeurs numériques. Il s'agit de généralisations de la recherche de facteurs dans un texte. L'alphabet est muni d'une relation d'ordre et plusieurs notions de motifs existent dans la littérature.

1 Notions de motifs

Order preserving matching Soient w et v deux mots de longueur n sur un alphabet totalement ordonné Σ. On parle d'équivalence d'ordre, notée $w \sim_o v$ si $w[i] < w[j]$ implique $v[i] < v[j]$ pour toute position $i, j \in \{1, \ldots, n\}$.

Par exemple : $(5, 2, 9, 4, 3) \sim_o (6, 1, 7, 5, 2)$

w et v sont donc équivalent si l'ordre relatif de leur valeur est le même.

Le problème de l'*order preserving matching* (OPM) est le fait de trouver dans un texte u, tous les facteurs v de u équivalents à un motif donné w.

Les arbres cartésiens Soit w un mot de longueur n sur un alphabet totalement ordonné Σ. L'arbre cartésien $CT(w)$ de w est un arbre binaire tel que :

- sa racine correspond à l'indice i de l'élément minimal de w. Afin que l'arbre associé à un mot soit unique, on décrète que s'il existe plusieurs occurrences de l'élément minimal, on choisit la position la plus à gauche dans le mot.
- le sous-arbre gauche de la racine est l'arbre cartésien du mot $w_1 \cdots w_{i-1}$.
- le sous-arbre droit de la racine est l'arbre cartésien du mot $w_{i+1} \cdots w_n$.

Il est possible de reconstruire une permutation à partir d'un arbre cartésien en effectuant un simple parcours préfixe de l'arbre. Les arbres cartésiens furent introduits par Vuillemin [1] au début des années 80 et des liens ont depuis été montrés avec les mots de Lyndon [2], les Range Minimum Queries [3] ou la construction d'arbres des suffixes en parallèle [4]. Récemment, Park et al. [5] ont introduit un nouveau type de motif généralisé sur les séquences, appelé

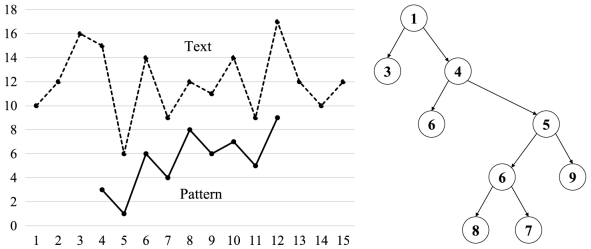

FIGURE 1 – Soit le texte $(10, 12, 16, 15, 6, 14, 9, 12, 11, 14, 9, 17, 12, 10, 12)$ et le motif $(3, 1, 6, 4, 8, 6, 7, 5, 9)$. Le facteur de longueur 8 commençant à la position 4 dans le texte, soit $(15, 6, 14, 9, 12, 11, 14, 9, 17)$, possède le même arbre cartésien (à droite) que le motif.

Cartesian Tree Matching (CTM). Le problème consiste à trouver tous les facteurs d'un texte v qui possèdent le même arbre cartésien qu'un motif donné w. Il s'agit d'une des généralisations du problème d'équivalence d'ordre. En effet, si $v \sim_o w$, alors $CT(v) = CT(w)$ (l'inverse n'étant pas nécessairement vrai). Park et al. ont mis en évidence une bijection entre les arbres cartésiens et une table des distance entre certains nœuds et leur parent. Une table leur permet d'obtenir des équivalents des algorithmes linéaires de recherche de Morris-Pratt (un seul motif) et d'Aho-Corasick (ensemble fini de motifs) pour le problème *CTM*. Des solutions plus efficaces en pratique ont été données dans [6, 7]. De plus, de nouveaux résultats sur le problème CTM ont été publiés [5, 8, 4] ces dernières années. Dans [9, 10], les auteurs se sont intéressés au calcul du plus long facteur Cartésien commun entre deux chaînes, c'est-à-dire des facteurs de longueurs maximales partageant le même arbre Cartésien, ce qui est utile pour découvrir des motifs intéressants où pour mesurer les similarités entre des séries temporelles.

2 Sujet de stage : approches probabilistes et motifs approchés

L'objectif de ce stage est de réfléchir à des notions de recherche approchée pour le problème CTM. Les distances de Hamming et de Levenshtein seront les premières à être considérées mais devront pour cela être définies formellement dans un tel contexte. Une version approchée du problème OPM a également été définie [11] et pourrait s'avérer utile pour définir une version équivalente pour le problème CTM. De plus, nous souhaitons avoir une approche probabiliste de ces problèmes. Le problème CTM peut être vu comme une version approchée du problème OPM, puisqu'il s'agit d'une généralisation. Nous souhaitons étudier pour un motif donné, la probabilité qu'un motif satisfasse le problème OPM à une position donnée sachant qu'il satisfait le problème CTM. L'étude de l'écart-type d'une telle valeur nous renseignera sur la disparité qui existe entre les différents motifs d'arbres Cartésiens et leur puissance d'expressivité.

Il s'agit d'une première approche, reliée à un sujet de thèse RIN en cours

de préparation, sur la recherche de motifs approchées dans les séquences ordonnées. L'analyse combatoire ou probabiliste joue ici un rôle double : mesurer à quel point une notion de motif diffère d'une autre, mais également mesurer l'efficacité moyenne des algorithmes associés. Le stage sera conjointement encadré par Thierry Lecroq (LITIS, Rouen) et Julien David (GREYC, Caen), dont les thématiques de recherches complémentaires ont permis la création de ce sujet. Le sujet mélange des aspects d'algorithmique du texte et des structures de données qui s'y rattache, soit l'un des domaines de recherche de Thierry Lecroq, ainsi que de la combinatoire et de l'analyse probabiliste des algorithmes, domaine de recherche de Julien David.

Références

[1] J. Vuillemin, "A unifying look at data structures," *Commun. ACM*, vol. 23, p. 229–239, apr 1980.

[2] M. Crochemore and L. M. Russo, "Cartesian and Lyndon trees," *Theoretical Computer Science*, vol. 806, pp. 1–9, 2020.

[3] E. Demaine, G. Landau, and O. Weimann, "On cartesian trees and range minimum queries," vol. 68, pp. 341–353, 01 2009.

[4] J. Shun and G. E. Blelloch, "A simple parallel cartesian tree algorithm and its application to parallel suffix tree construction," *ACM Trans. Parallel Comput.*, vol. 1, oct 2014.

[5] S. Park, A. Amir, G. Landau, and K. Park, "Cartesian tree matching and indexing," in *30th Annual Symposium on Combinatorial Pattern Matching, CPM 2019* (N. Pisanti and S. Pissis, eds.), Leibniz International Proceedings in Informatics, LIPIcs, (Germany), pp. 16 :1–16 :14, Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, June 2019.

[6] S. Song, G. Gu, C. Ryu, S. Faro, T. Lecroq, and K. Park, "Fast cartesian tree matching," (Segovia, Spain), pp. 124–137, 2019.

[7] S. Song, G. Gu, C. Ryu, S. Faro, T. Lecroq, and K. Park, "Fast algorithms for single and multiple pattern cartesian tree matching," *Theor. Comput. Sci.*, vol. 849, pp. 47–63, 2021.

[8] S. G. Park, M. Bataa, A. Amir, G. M. Landau, and K. Park, "Finding patterns and periods in cartesian tree matching," *Theoretical Computer Science*, vol. 845, pp. 181–197, 2020.

[9] S. Faro, T. Lecroq, and K. Park, "Fast practical computation of the longest common cartesian substrings of two strings," in *Stringology*, PSC, (Prague, Czech Republic), pp. 48–60, 2020.

[10] S. Faro, T. Lecroq, K. Park, and S. Scafiti, "On the longest common cartesian substring problem," *The Computer Journal*, 01 2022.

[11] P. Gawrychowski and P. Uznanski, "Order-preserving pattern matching with k mismatches," vol. 8486, 09 2013.