MITRO210 : Automates et données structurées

Feuille d'exercices 3

Antoine Amarilli

1 Algorithme d'Aho-Corasick

Dans cet exercice, on cherche à généraliser l'algorithme de Knuth-Morris-Pratt, vu en cours, pour couvrir le cas où il y a plusieurs chaînes de caractères à rechercher.

Dans les questions de 0 à 2, on travaillera avec l'ensemble de mots $U_0 = \{ab, bab, bca, caa\}$.

Question 0. On cherche dans un premier temps à construire un automate qui accepte *exactement* le langage U_0. En proposer un, en s'inspirant de l'idée des tries.

Question 1. On cherche à présent à construire un automate qui accepte tous les mots qui *se terminent par* U_0. Modifier l'automate précédent pour qu'il remplisse cette fonction. On utilisera des transitions d'échec.

Question 2. Expliquer comment on peut utiliser l'automate de la question 1 pour trouver, étant donné un mot v, quelle est à chaque position de v le plus long suffixe se finissant à cet endroit qui soit un préfixe d'un mot de U_0. Quelle est la complexité de cet algorithme ?

Question 3. En généralisant les questions 0 et 1, proposer un algorithme général pour la tâche suivante : étant donné un ensemble fini de mots U, construire un automate qui reconnaît les mots qui se terminent par U_0. Quelle est sa complexité ?

Question 4. En déduire un algorithme pour la tâche suivante : étant donné un ensemble fini de mots U, et un mot v, trouver à chaque position de v quel est le plus long suffixe se finissant à cet endroit qui soit un préfixe d'un mot de U. Quelle est la complexité de cet algorithme ?

Question 5. En déduire enfin un algorithme pour la tâche suivante : étant donné un ensemble fini de mots U, et un mot v, trouver toutes les occurrences d'un mot de U dans v. Quelle est la complexité de cet algorithme ?