Recherche rapide

Philippe Biondi
<biondi@cartel-securite.fr>

12 mars 2002

Résumé

Cet article traite de queleques algorithmes de recherche rapide d'un ensemble de signatures. Les applications peuvent être diverses. On peut citer par exemple les systèmes de détection d'intrusion ou les antivirus.

Table des matières

1 Énoncé du problème

Soit Σ un alphabet. Soit . l'opérateur de concaténation. Soit $(\Sigma^*, .)$ le monoïde engendré par $(\Sigma, .)$. On notera ε son élément neutre (mot vide). Σ^* est isomorphe à l'ensemble des suites finies (mots) sur Σ et nous utiliserons indistinctement le vocabulaire de l'un ou de l'autre.

Soit t un élément de Σ^*. Soit $S \subset \Sigma^*$ un ensemble fini de mots. S sera appelé *base de signatures*.

Le problème consiste à déterminer le plus rapidement possible si un des éléments de S est une suite extraite de t. Autrement dit, étant donnés S et t,

$$\exists?(s, d, f) \in S \times \Sigma^* \times \Sigma^* \quad | \quad t = d.s.f$$

2 Résolution par des algorithmes de type Boyer-Moore

3 Résolution par un automate fini déterministe

3.1 Algorithme 1

L'algorithme 1 consiste à construire un automate fini déterministe capable de reconnaître toutes les signatures de la base en un seul passage. La construction se déroule en 3 étapes.

3.1.1 Construction de l'automate

Construction du squelette. La première étape consiste à construire le squelette de l'automate. Soit l'automate $\mathcal{A}_0 = (E, T, i, F)$ où E est l'ensemble des états de l'automate, $T \subset E \times \Sigma \times E$ est l'ensemble de ses transitions, $i \in E$ est l'état initial et $T \subset E$ est l'ensemble des états terminaux.

$$\begin{aligned} E &= \{x \in \Sigma^* \quad | \quad \exists y \in \Sigma^* \ x.y \in S\} \\ T &= \{(x,t,y) \in E \times \Sigma \times E \quad | \quad x.t = y\} \\ i &= \varepsilon \\ T &= S \end{aligned} \tag{1}$$

Construction de l'arbre inversé. Il s'agit ensuite de construire un arbre référençant tous les nœuds de G par la fin. Soit donc l'arbre $A = (N, D)$ tel que

$$\begin{aligned} N &= \{x \in \Sigma^* \quad | \quad \exists y \in \Sigma^* \ x.y \in E\} \\ D &= \{(x,y) \in N^2 \quad | \quad \exists t \in \Sigma \ x.t = y\} \end{aligned} \tag{2}$$

Finition de l'automate. Maintenant que nous avons son squelette $\mathcal{A}$, nous allons compléter les transitions manquantes.

Pour chaque $e \in E$, on considère dans l'arbre A la chaîne $(x_0, \ldots, x_m)$ telle que $\forall i \quad 0 \leq i < m \Longrightarrow (x_i, x_{i+1}) \in D$ et que $x_0 = \varepsilon$ et $x_m = e$. On ne veut garder que les éléments de E. Soit donc $(y_0, \ldots, y_n)$ telle que $y_0 = x_0$ et que $\forall i \in \{0, \ldots, m\}$ $\exists (j,k)$ tel que

$$\begin{aligned} &0 \leq j < k \leq m \\ &y_i = x_j \in E \\ &y_{i+1} = x_k \in E \\ &\forall l j < l < k \Longrightarrow x_l \notin E \end{aligned}$$

Soit $X_i = \mathrm{proj}_{2,3}\,(T \bigcap (\{y_i\} \times \Sigma \times E))$ l'ensemble des transitions partant de l'état y_i. Soit $Y_i = \mathrm{proj}_1(X_i)$. Soit $Z_i = Y_i - Y_{i+1} - \ldots - Y_n$. Soit $U_e = \bigcup_{j=0}^{n} (X_j \bigcap (Z_j \times E))$. S'il existe y_i tel que $y_i \in F$ alors $F_e = e$, sinon, $F_e =$

Une fois U_e et F_e calculés pour tout e. On pose $T' = \bigcup_{e \in E} (\{e\} \times T_e)$ et $F' = \bigcup_{e \in E} F_e$. On considère alors l'automate $\mathcal{A} = (E, T', i, F')$.

3.1.2 Utilisation

Théorème 3.1 :
L'automate $\mathcal{A}$ reconnaît tous les éléments de S et ne reconnaît que ceux-ci.

Démonstration. Ben c'est parti.. $\square$

4 Résolution par un automate fini non-déterminisite