Présenter une liste de dates de manière lisible : complexité et algorithme

Guillaume Pinot

Canal TP
20 rue Hector Malot
75012 Paris, France
guillaume.pinot@canaltp.fr

ROADEF 2015

Présentation de Canal TP

Canal TP :

- éditeur logiciel fournissant expertise et solutions ;
- centré sur l'information voyageurs.

navitia :

- un moteur disponible en logiciel libre `https://github.com/CanalTP/navitia` ;
- un *web service* `http://navitia.io` ;
- présentation de l'offre de transport : lignes, arrêts, horaires ;
- autocompletion et géolocalisation ;
- calcul d'itinéraires multimodaux porte-à-porte.

Problématique

Calendrier

Lu	Ma	Me	Je	Ve	Sa	Di
27	**28**	**29**	**30**	1	2	3
4	**5**	**6**	**7**	8	9	10
11	**12**	**13**	14	**15**	16	17
18	**19**	**20**	**21**	**22**	23	24
25	**26**	**27**	**28**	**29**	**30**	31

Sortie

1111000111100011101001111100011110

Et pour un humain ?

Table des matières

Table des matières

Description du problème

Calendrier

Lu	Ma	Me	Je	Ve	Sa	Di
27	**28**	**29**	**30**	*1*	2	3
4	**5**	**6**	**7**	*8*	9	10
11	**12**	**13**	*14*	**15**	16	17
18	**19**	**20**	**21**	**22**	23	24
25	**26**	**27**	**28**	**29**	***30***	31

Sortie humaine

Du Lundi au Vendredi	un rythme hebdomadaire
du 27 avril au 31 mai 2015	une période de validité
sauf les 1, 8, 14 et 25 mai,	une liste de date exclues
avec en plus le 30 mai.	une liste de date incluses
Objectif :	minimiser le nombre d'exceptions

Description du problème

Calendrier

Lu	Ma	Me	Je	Ve	Sa	Di
27	**28**	**29**	**30**	*1*	2	3
4	**5**	**6**	**7**	*8*	9	10
11	**12**	**13**	*14*	**15**	16	17
18	**19**	**20**	**21**	**22**	23	24
25	**26**	**27**	**28**	**29**	***30***	31

Sortie humaine

Du Lundi au Vendredi	un rythme hebdomadaire
du 27 avril au 31 mai 2015	une période de validité
sauf les 1, 8, 14 et 25 mai,	une liste de date exclues
avec en plus le 30 mai.	une liste de date incluses
Objectif :	minimiser le nombre d'exceptions

Description du problème

Calendrier

Lu	Ma	Me	Je	Ve	Sa	Di
27	**28**	**29**	**30**	*1*	2	3
4	**5**	**6**	**7**	*8*	9	10
11	**12**	**13**	*14*	**15**	16	17
18	**19**	**20**	**21**	**22**	23	24
25	**26**	**27**	**28**	**29**	***30***	31

Sortie humaine

Du Lundi au Vendredi	un rythme hebdomadaire
du 27 avril au 31 mai 2015	une période de validité
sauf les 1, 8, 14 et 25 mai,	une liste de date exclues
avec en plus le 30 mai.	une liste de date incluses
Objectif :	minimiser le nombre d'exceptions

Données du problème

Calendrier

Lu	Ma	Me	Je	Ve	Sa	Di
27	**28**	**29**	**30**	1	2	3
4	**5**	**6**	**7**	8	9	10
11	**12**	**13**	14	**15**	16	17
18	**19**	**20**	**21**	**22**	23	24
25	**26**	**27**	**28**	**29**	**30**	31

Données du problème

1111000111100011101001111000111110

Les rythmes hebdomadaires de chaque semaine.

Données du problème

Calendrier

Lu	Ma	Me	Je	Ve	Sa	Di
27	**28**	**29**	**30**	1	2	3
4	**5**	**6**	**7**	8	9	10
11	**12**	**13**	14	**15**	16	17
18	**19**	**20**	**21**	**22**	23	24
25	**26**	**27**	**28**	**29**	**30**	31

Données du problème

$$\mathcal{S} = 1111000, 1111000, 1110100, 1111100, 0111110$$

Les rythmes hebdomadaires de chaque semaine.

Données du problème

Calendrier

Lu	Ma	Me	Je	Ve	Sa	Di
27	**28**	**29**	**30**	1	2	3
4	**5**	**6**	**7**	8	9	10
11	**12**	**13**	14	**15**	16	17
18	**19**	**20**	**21**	**22**	23	24
25	**26**	**27**	**28**	**29**	**30**	31

Données du problème

$$\mathcal{S} = 1111000, 1111000, 1110100, 1111100, 0111110$$

Les rythmes hebdomadaires de chaque semaine.

Distance de Hamming

Comment exprimer un rythme avec un autre et des exceptions ?

Compter les exceptions

Du Mardi au Samedi	$s_1 = 0111110$
Du Lundi au Vendredi	$s_2 = 1111100$
Distance de Hamming [Hamming, 1950]	$d(s_1, s_2) = 2$

Distance de Hamming

Comment exprimer un rythme avec un autre et des exceptions ?

Compter les exceptions

Du Mardi au Samedi	$s_1 = 0111110$
Du Lundi au Vendredi	$s_2 = 1111100$
Distance de Hamming [Hamming, 1950]	$d(s_1, s_2) = 2$

Distance de Hamming

Comment exprimer un rythme avec un autre et des exceptions ?

Compter les exceptions

Du Mardi au Samedi	$s_1 = 0111110$
Du Lundi au Vendredi	$s_2 = 1111100$
Distance de Hamming [Hamming, 1950]	$d(s_1, s_2) = 2$

canal tp

Modèle mathématique

Le meilleur rythme hebdomadaire

Soit $\mathcal{S} = s_1, s_2, \ldots, s_n$, n chaînes binaires de longueur 7.
Soit $d_{s_i}^{\Sigma} = \sum_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le meilleur rythme hebdomadaire est représenté par la chaîne s tel que d_s^{Σ} soit minimale.

Notre exemple

$$\mathcal{S} = 1111{\color{red}0}00, 1111{\color{red}0}00, 111{\color{red}0}100, 1111100, {\color{red}0}1111{\color{red}1}0$$

Pour le rythme « du Lundi au Vendredi », $s_i = 1111100$:

$$d_{s_i}^{\Sigma} = \sum_{s_j \in \mathcal{S}} d(s_i, s_j) = 1 + 1 + 1 + 0 + 2 = 5$$

Table des matières

Le *closest string problem*

Le *closest string problem*

Soit $\mathcal{S} = s_1, s_2, \ldots, s_n$, n chaînes sur l'alphabet Σ de longueur m.
Soit $d_{s_i}^{\max} = \max_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le *closest string problem* a pour solution la chaîne s tel que $d_s^{\max}$ soit minimale.

Complexité :

- NP-difficile [Lanctot et al., 2003]
- $O(nm + nd_s^{\max}(16|\Sigma|)^{d_s^{\max}})$ [Ma and Sun, 2008] (exponentielle sur m)

Comparaison des deux problèmes

Le meilleur rythme hebdomadaire

Soit $\mathcal{S} = s_1, s_2, \dots, s_n$, n chaînes binaires de longueur 7.
Soit $d_{s_i}^{\Sigma} = \sum_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le meilleur rythme hebdomadaire est représenté par la chaîne s tel que d_s^{Σ} soit minimale.

Le *closest string problem*

Soit $\mathcal{S} = s_1, s_2, \dots, s_n$, n chaînes sur l'alphabet Σ de longueur m.
Soit $d_{s_i}^{\max} = \max_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le *closest string problem* a pour solution la chaîne s tel que $d_s^{\max}$ soit minimale.

Comme le *closest string problem* est exponentiel sur m, nous pouvons espérer trouver un algorithme polynomial à notre problème.

Comparaison des deux problèmes

Le meilleur rythme hebdomadaire

Soit $\mathcal{S} = s_1, s_2, \ldots, s_n$, n chaînes binaires de longueur 7.
Soit $d_{s_i}^{\Sigma} = \sum_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le meilleur rythme hebdomadaire est représenté par la chaîne s tel que d_s^{Σ} soit minimale.

Le *closest string problem*

Soit $\mathcal{S} = s_1, s_2, \ldots, s_n$, n chaînes sur l'alphabet Σ de longueur m.
Soit $d_{s_i}^{\max} = \max_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le *closest string problem* a pour solution la chaîne s tel que $d_s^{\max}$ soit minimale.

Comme le *closest string problem* est exponentiel sur m, nous pouvons espérer trouver un algorithme polynomial à notre problème.

Comparaison des deux problèmes

Le meilleur rythme hebdomadaire

Soit $\mathcal{S} = s_1, s_2, \ldots, s_n$, n chaînes binaires de longueur 7.
Soit $d^{\Sigma}_{s_i} = \sum_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le meilleur rythme hebdomadaire est représenté par la chaîne s tel que d^{Σ}_{s} soit minimale.

Le *closest string problem*

Soit $\mathcal{S} = s_1, s_2, \ldots, s_n$, n chaînes sur l'alphabet Σ de longueur m.
Soit $d^{\max}_{s_i} = \max_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le *closest string problem* a pour solution la chaîne s tel que $d^{\max}_{s}$ soit minimale.

Comme le *closest string problem* est exponentiel sur m, nous pouvons espérer trouver un algorithme polynomial à notre problème.

Comparaison des deux problèmes

Le meilleur rythme hebdomadaire

Soit $\mathcal{S} = s_1, s_2, \ldots, s_n$, n chaînes binaires de longueur 7.
Soit $d_{s_i}^{\Sigma} = \sum_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le meilleur rythme hebdomadaire est représenté par la chaîne s tel que d_s^{Σ} soit minimale.

Le *closest string problem*

Soit $\mathcal{S} = s_1, s_2, \ldots, s_n$, n chaînes sur l'alphabet Σ de longueur m.
Soit $d_{s_i}^{\max} = \max_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le *closest string problem* a pour solution la chaîne s tel que $d_s^{\max}$ soit minimale.

Comme le *closest string problem* est exponentiel sur m, nous pouvons espérer trouver un algorithme polynomial à notre problème.

Comparaison des deux problèmes

Le meilleur rythme hebdomadaire

Soit $\mathcal{S} = s_1, s_2, \dots, s_n$, n chaînes binaires de longueur 7.
Soit $d^{\Sigma}_{s_i} = \sum_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le meilleur rythme hebdomadaire est représenté par la chaîne s tel que d^{Σ}_{s} soit minimale.

Le *closest string problem*

Soit $\mathcal{S} = s_1, s_2, \dots, s_n$, n chaînes sur l'alphabet Σ de longueur m.
Soit $d^{\max}_{s_i} = \max_{s_j \in \mathcal{S}} d(s_i, s_j)$.
Le *closest string problem* a pour solution la chaîne s tel que $d^{\max}_{s}$ soit minimale.

Comme le *closest string problem* est exponentiel sur m, nous pouvons espérer trouver un algorithme polynomial à notre problème.

Table des matières

Énumération exhaustive

Soit $\mathcal{D}$ l'ensemble des chaînes binaires de longueur 7.

$s \leftarrow 0000000$

$d_s^\Sigma \leftarrow \infty$

$\forall s_i \in \mathcal{D}$: $|\Sigma|^m = 2^7 = 128$ *fois*

- calculer $d_{s_i}^\Sigma = \sum_{s_j \in \mathcal{S}} d(s_i, s_j)$ $O(mn) = O(7n) = O(n)$
- si $d_{s_i}^\Sigma < d_s^\Sigma$:
  - $d_s^\Sigma \leftarrow d_{s_i}^\Sigma$
  - $s \leftarrow s_i$

Résultat : s est le meilleur rythme hebdomadaire, avec d_s^Σ exceptions.

Complexité : $O(mn|\Sigma|^m) = O(128 \times 7n) = O(n)$

canal tp

Énumération exhaustive

Soit $\mathcal{D}$ l'ensemble des chaînes binaires de longueur 7.
$s \leftarrow 0000000$
$d_s^\Sigma \leftarrow \infty$
$\forall s_i \in \mathcal{D}$: $\qquad |\Sigma|^m = 2^7 = 128$ *fois*

- calculer $d_{s_i}^\Sigma = \sum_{s_j \in \mathcal{S}} d(s_i, s_j)$ $\qquad O(mn) = O(7n) = O(n)$
- si $d_{s_i}^\Sigma < d_s^\Sigma$:
  - $d_s^\Sigma \leftarrow d_{s_i}^\Sigma$
  - $s \leftarrow s_i$

Résultat : s est le meilleur rythme hebdomadaire, avec d_s^Σ exceptions.

Complexité : $O(mn|\Sigma|^m) = O(128 \times 7n) = O(n)$

Modélisation matricielle

$$\mathcal{S} = 101, 110, 110, 011, \ldots = 4\times 011, 3\times 101, 5\times 110 = 4s_3, 3s_5, 5s_6$$

$$\begin{pmatrix} 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \\ 1 & 0 & 2 & 1 & 2 & 1 & 3 & 2 \\ 1 & 2 & 0 & 1 & 2 & 3 & 1 & 2 \\ 2 & 1 & 1 & 0 & 3 & 2 & 2 & 1 \\ 1 & 2 & 2 & 3 & 0 & 1 & 1 & 2 \\ 2 & 1 & 3 & 2 & 1 & 0 & 2 & 1 \\ 2 & 3 & 1 & 2 & 1 & 2 & 0 & 1 \\ 3 & 2 & 2 & 1 & 2 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 3 \\ 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 24 \\ 22 \\ 18 \\ 16 \\ 20 \\ 18 \\ 14 \\ 12 \end{pmatrix}$$

$s_0 = 000$
$s_1 = 001$
$s_2 = 010$
$s_3 = 011$
$s_4 = 100$
$s_5 = 101$
$s_6 = 110$
$s_7 = 111$

$$d_s^\Sigma = d_{s_7}^\Sigma = 12, s = s_7 = 111$$

Modélisation matricielle

$$\mathcal{S} = 101, 110, 110, 011, \ldots = 4 \times 011, 3 \times 101, 5 \times 110 = 4s_3, 3s_5, 5s_6$$

$$\begin{pmatrix} 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \\ 1 & 0 & 2 & 1 & 2 & 1 & 3 & 2 \\ 1 & 2 & 0 & 1 & 2 & 3 & 1 & 2 \\ 2 & 1 & 1 & 0 & 3 & 2 & 2 & 1 \\ 1 & 2 & 2 & 3 & 0 & 1 & 1 & 2 \\ 2 & 1 & 3 & 2 & 1 & 0 & 2 & 1 \\ 2 & 3 & 1 & 2 & 1 & 2 & 0 & 1 \\ 3 & 2 & 2 & 1 & 2 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 3 \\ 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 24 \\ 22 \\ 18 \\ 16 \\ 20 \\ 18 \\ 14 \\ 12 \end{pmatrix} \qquad \begin{array}{l} s_0 = 000 \\ s_1 = 001 \\ s_2 = 010 \\ s_3 = 011 \\ s_4 = 100 \\ s_5 = 101 \\ s_6 = 110 \\ s_7 = 111 \end{array}$$

$$d_s^{\Sigma} = d_{s_7}^{\Sigma} = 12, s = s_7 = 111$$

Modélisation matricielle

$$\mathcal{S} = 101, 110, 110, 011, \ldots = 4 \times 011, 3 \times 101, 5 \times 110 = 4s_3, 3s_5, 5s_6$$

$$\begin{pmatrix} 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \\ 1 & 0 & 2 & 1 & 2 & 1 & 3 & 2 \\ 1 & 2 & 0 & 1 & 2 & 3 & 1 & 2 \\ 2 & 1 & 1 & 0 & 3 & 2 & 2 & 1 \\ 1 & 2 & 2 & 3 & 0 & 1 & 1 & 2 \\ 2 & 1 & 3 & 2 & 1 & 0 & 2 & 1 \\ 2 & 3 & 1 & 2 & 1 & 2 & 0 & 1 \\ 3 & 2 & 2 & 1 & 2 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 3 \\ 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 24 \\ 22 \\ 18 \\ 16 \\ 20 \\ 18 \\ 14 \\ 12 \end{pmatrix} \qquad \begin{aligned} s_0 &= 000 \\ s_1 &= 001 \\ s_2 &= 010 \\ s_3 &= 011 \\ s_4 &= 100 \\ s_5 &= 101 \\ s_6 &= 110 \\ s_7 &= 111 \end{aligned}$$

$$d_s^{\Sigma} = d_{s_7}^{\Sigma} = 12, s = s_7 = 111$$

Modélisation matricielle

$$\mathcal{S} = 101, 110, 110, 011, \ldots = 4 \times 011, 3 \times 101, 5 \times 110 = 4s_3, 3s_5, 5s_6$$

$$\begin{pmatrix} 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \\ 1 & 0 & 2 & 1 & 2 & 1 & 3 & 2 \\ 1 & 2 & 0 & 1 & 2 & 3 & 1 & 2 \\ 2 & 1 & 1 & 0 & 3 & 2 & 2 & 1 \\ 1 & 2 & 2 & 3 & 0 & 1 & 1 & 2 \\ 2 & 1 & 3 & 2 & 1 & 0 & 2 & 1 \\ 2 & 3 & 1 & 2 & 1 & 2 & 0 & 1 \\ 3 & 2 & 2 & 1 & 2 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 3 \\ 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 24 \\ 22 \\ 18 \\ 16 \\ 20 \\ 18 \\ 14 \\ 12 \end{pmatrix} \qquad \begin{aligned} s_0 &= 000 \\ s_1 &= 001 \\ s_2 &= 010 \\ s_3 &= 011 \\ s_4 &= 100 \\ s_5 &= 101 \\ s_6 &= 110 \\ s_7 &= 111 \end{aligned}$$

$$d_s^{\Sigma} = d_{s_7}^{\Sigma} = 12, s = s_7 = 111$$

Modélisation matricielle

$$\mathcal{S} = 101, 110, 110, 011, \ldots = 4 \times 011, 3 \times 101, 5 \times 110 = 4s_3, 3s_5, 5s_6$$

$$\begin{pmatrix} 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \\ 1 & 0 & 2 & 1 & 2 & 1 & 3 & 2 \\ 1 & 2 & 0 & 1 & 2 & 3 & 1 & 2 \\ 2 & 1 & 1 & 0 & 3 & 2 & 2 & 1 \\ 1 & 2 & 2 & 3 & 0 & 1 & 1 & 2 \\ 2 & 1 & 3 & 2 & 1 & 0 & 2 & 1 \\ 2 & 3 & 1 & 2 & 1 & 2 & 0 & 1 \\ 3 & 2 & 2 & 1 & 2 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 3 \\ 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 24 \\ 22 \\ 18 \\ 16 \\ 20 \\ 18 \\ 14 \\ 12 \end{pmatrix} \qquad \begin{aligned} s_0 &= 000 \\ s_1 &= 001 \\ s_2 &= 010 \\ s_3 &= 011 \\ s_4 &= 100 \\ s_5 &= 101 \\ s_6 &= 110 \\ s_7 &= 111 \end{aligned}$$

$$d_s^{\Sigma} = d_{s_7}^{\Sigma} = 12, s = s_7 = 111$$

Table des matières

Conclusions et perspectives

Conclusions :

- présentation d'un petit problème pratique ;
- problème proche du *closest string problem* qui est NP-difficile ;
- le problème est malgré tout polynômial.

Perspectives :

- semaines partielles sur les bords ;
- plusieurs rythmes.

Conclusions et perspectives

Conclusions :

- présentation d'un petit problème pratique ;
- problème proche du *closest string problem* qui est NP-difficile ;
- le problème est malgré tout polynômial.

Perspectives :

- semaines partielles sur les bords ;
- plusieurs rythmes.

Présenter une liste de dates de manière lisible : complexité et algorithme

Guillaume Pinot

Canal TP
20 rue Hector Malot
75012 Paris, France
guillaume.pinot@canaltp.fr

ROADEF 2015

Bibliographie I

Hamming, R. (1950).
Error detecting and error correcting codes.
Bell System Technical Journal, 26(2) :147–160.

Lanctot, J. K., Lia, M., Mab, B., Wangc, S., and Zhang, L. (2003).
Distinguishing string selection problems.
Information and Computation, 185 :41–55.

Ma, B. and Sun, X. (2008).
More efficient algorithms for closest string and substring problems.
In *Research in Computational Molecular Biology*, volume 4955, pages 396–409. Springer.

canal tp