



Searching for Dependencies at Multiple Abstraction Levels

Toon Calders, Raymond T. Ng et Jef Wijsen

Matthieu Objois et Pierre Senellart

1. Introduction

✓ Cadre

2. Motivations

3. Problème RUDMINE

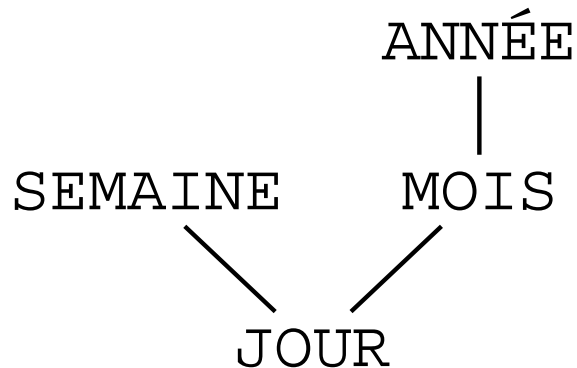
4. Algorithme

5. Conclusion

Cadre : hiérarchies de généralisation

- ⑥ Bases de données relationnelles
- ⑥ Hiérarchie sur les valeurs des attributs (*niveaux*)

EXEMPLE



1. Introduction
✓ Dépendance roll-up
2. Motivations
3. Problème RUDMINE
4. Algorithme
5. Conclusion

Dépendance roll-up (DRU)

- ⑥ Extension de la notion de dépendance fonctionnelle
- ⑥ EXEMPLE

$(Ville, RÉGION) \rightarrow (Taux_Natalité, POURMILLE)$

Cette DRU est vérifiée



Au sein d'une même région, toutes les villes ont le même taux de natalité arrondi au 1‰ le plus proche.

1. Introduction
✓ Plan

- 2. Motivations
- 3. Problème RUDMINE
- 4. Algorithme
- 5. Conclusion

Plan de l'exposé

1. Introduction

- 2. Motivations
 - ✓ OLAP
 - ✓ Élimination de redondance
 - ✓ Prédiction de données

- 3. RUDMINE
 - ✓ Support
 - ✓ Confiance
 - ✓ Gen-schéma

- 4. Algorithme
 - ✓ Principe
 - ✓ Simulation
 - ✓ Performances

5. Conclusion

1. Introduction
2. Motivations
✓ OLAP
3. Problème RUDMINE
4. Algorithme
5. Conclusion

OnLine Analytic Processing

⑥ Déterminer la granularité intéressante pour OLAP

⑥ EXEMPLE

- △ $(Ville, RÉGION) \rightarrow (Taux_Natalité, POURMILLE)$
vérifiée
- △ $(Ville, PAYS) \rightarrow (Taux_Natalité, POURMILLE)$
non vérifiée

⇒ données intéressantes : taux de natalité par *région*

1. Introduction

2. Motivations

✓ Élimination de redondance

3. Problème RUDMINE

4. Algorithme

5. Conclusion

Élimination de redondance

⑥ Réduction du nombre de données

EXEMPLE

$(\text{Horaire}, \text{HEURE}), (\text{Station}, \text{RÉGION}) \rightarrow (\text{Temp}^\circ, \text{ENTIER})$
vérifiée

\implies on peut ne garder que les relevés par *heure* et par *région*

⑥ Aide à la conception de schémas

1. Introduction
2. Motivations
 - ✓ Prédiction de données
3. Problème RUDMINE
4. Algorithme
5. Conclusion

Prédiction de données

- ⑥ Découverte de DRU \Rightarrow possibilité de :
 - △ compléter des données manquantes
 - △ prévoir de nouvelles données
- ⑥ Analogie aux techniques classiques en Fouille de Données (règles d'association...)

1. Introduction
2. Motivations
3. Problème RUDMINE
✓ Support
4. Algorithme
5. Conclusion

Support d'une DRU

- ⑥ Caractérise le nombre de cas dans lesquels la règle peut s'appliquer

- ⑥ EXEMPLE

$(Ville, RÉGION) \rightarrow (Taux_Natalité, POURMILLE)$

support de cette DRU : proportion de paires de n-uplets de la base ayant des villes de la même région

1. Introduction
2. Motivations
3. Problème RUDMINE
✓ Confiance
4. Algorithme
5. Conclusion

Confiance d'une DRU

⑥ Caractérise le nombre de cas dans lesquels la règle est vérifiée

⑥ EXEMPLE

$(Ville, RÉGION) \rightarrow (Taux_Natalité, POURMILLE)$

confiance de cette DRU : proportion de paires de n-uplets de la base ayant le même taux de natalité arrondi au 1‰, parmi ceux ayant des villes de la même région

Schéma de généralisation (Gen-schéma)

1. Introduction
2. Motivations
3. Problème RUDMINE
✓ Gen-schéma
4. Algorithme
5. Conclusion

- ⑥ Un gen-schéma d'une relation est un ensemble de couples $(attribut, niveau)$ non redondant.
- ⑥ On note $G \trianglelefteq H$ si le niveau de chacun des attributs dans H est plus grand que celui dans G .
- ⑥ EXEMPLE

2 gen-schémas de $\{(Date, JOUR), (Produit, PRODUIT)\}$:

$$\{(Date, MOIS), (Produit, PRODUIT)\} \trianglelefteq \{(Date, ANNÉE)\}$$

1. Introduction
2. Motivations
3. Problème RUDMINE
4. Algorithme
✓ Principe
5. Conclusion

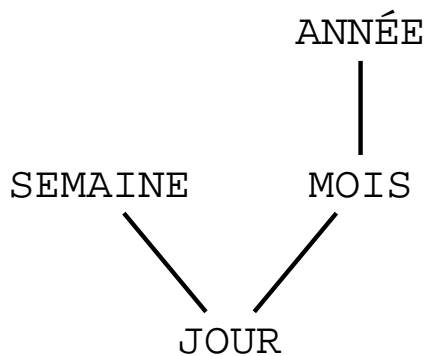
Principe de l'algorithme

- ⑥ Parcours de l'ensemble des gen-schémas (qui forment potentiellement la partie gauche de la DRU), du plus général au plus spécifique
- ⑥ À chaque étape (*strate*), génération des spécialisations immédiates des gen-schémas actuels
- ⑥ Test du support et de la confiance des candidats, pour retenir ceux qui dépassent les seuils fixés
- ⑥ Élagage possible

1. Introduction
2. Motivations
3. Problème RUDMINE
4. Algorithme
✓ Simulation
5. Conclusion

Simulation de l'algorithme

DRU de la forme : ? \rightarrow (Vente, k€) ($s_{seuil} = 0,25$; $c_{seuil} = 0,7$)

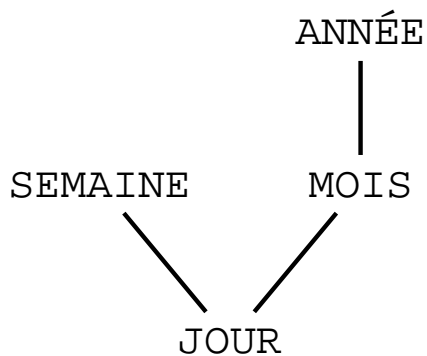


1. Introduction
2. Motivations
3. Problème RUDMINE
4. **Algorithme**
✓ **Simulation**
5. Conclusion

Simulation de l'algorithme

DRU de la forme : $? \rightarrow (Vente, k\text{€})$ ($s_{seuil} = 0,25$; $c_{seuil} = 0,7$)

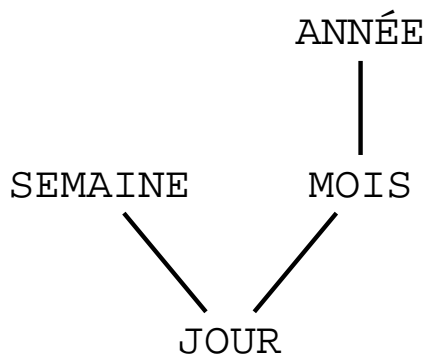
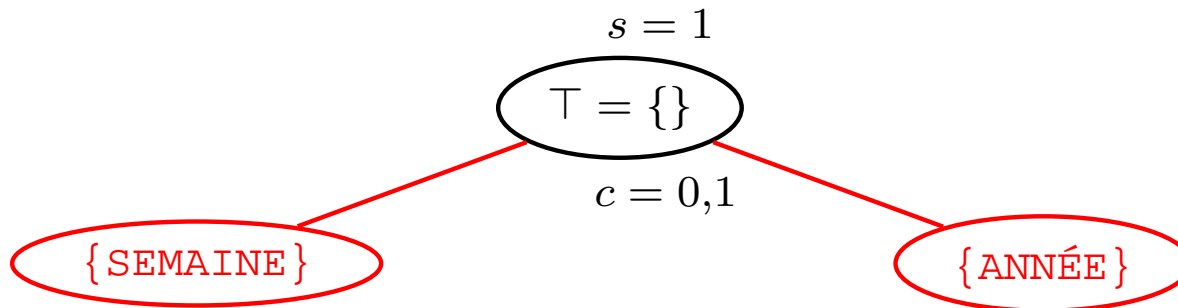
$$\begin{array}{c} s = 1 \\ \text{T} = \{\} \\ c = 0,1 \end{array}$$



1. Introduction
2. Motivations
3. Problème RUDMINE
4. **Algorithme**
✓ Simulation
5. Conclusion

Simulation de l'algorithme

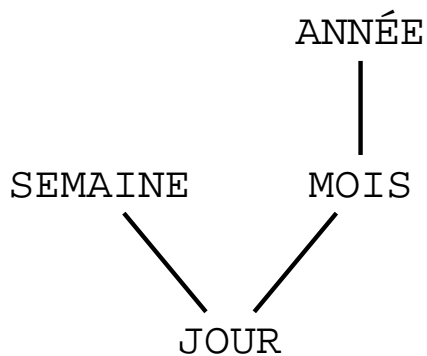
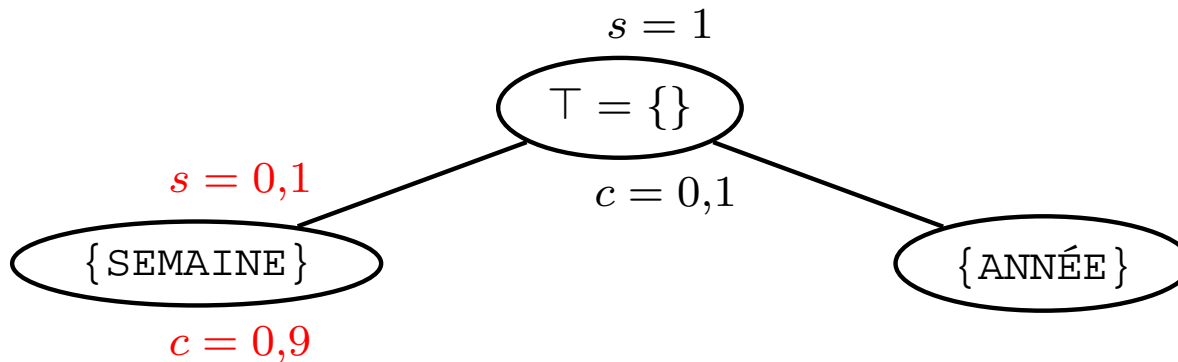
DRU de la forme : $? \rightarrow (Vente, k\text{€})$ ($s_{seuil} = 0,25$; $c_{seuil} = 0,7$)



1. Introduction
2. Motivations
3. Problème RUDMINE
4. **Algorithme**
✓ Simulation
5. Conclusion

Simulation de l'algorithme

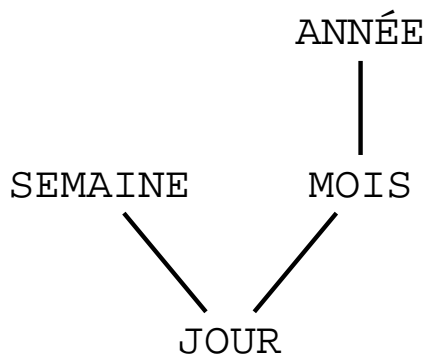
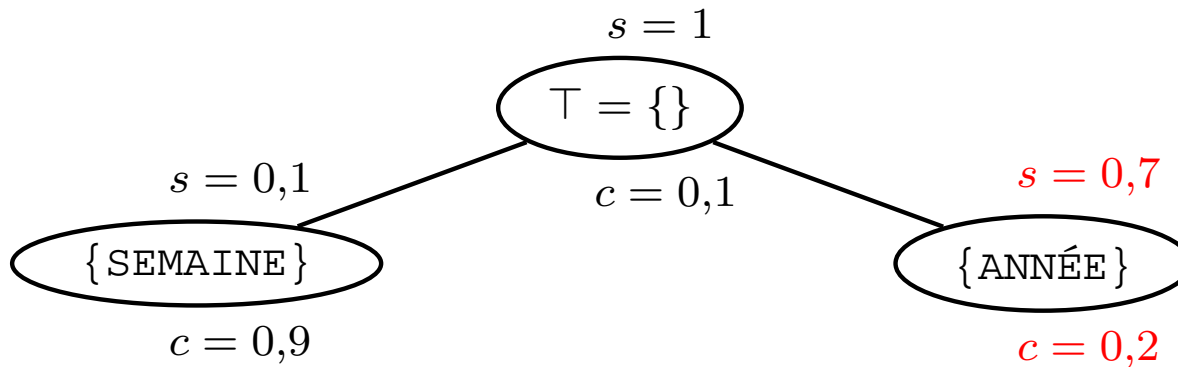
DRU de la forme : $? \rightarrow (Vente, k\text{€})$ ($s_{seuil} = 0,25$; $c_{seuil} = 0,7$)



1. Introduction
2. Motivations
3. Problème RUDMINE
4. **Algorithme**
✓ Simulation
5. Conclusion

Simulation de l'algorithme

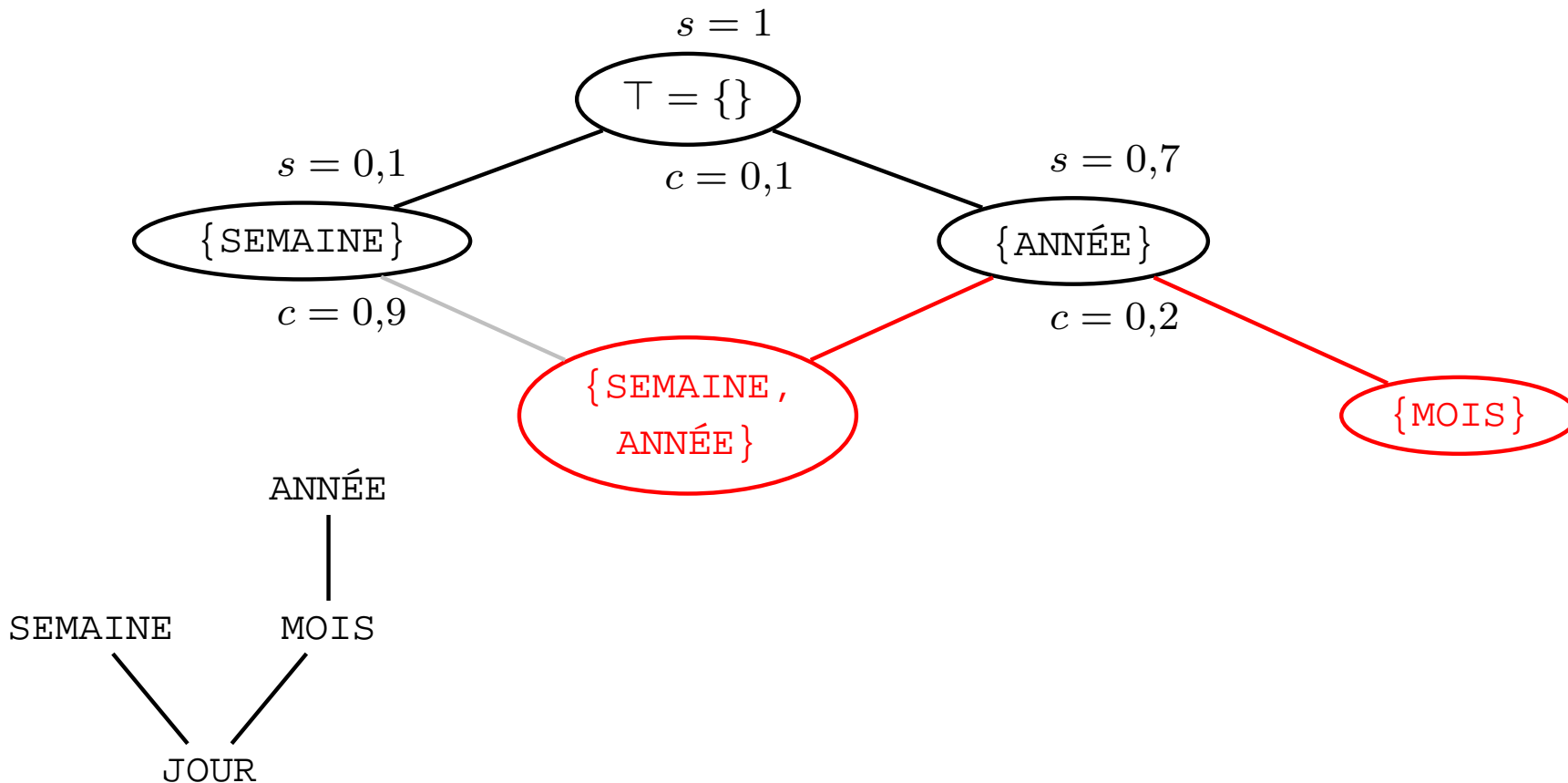
DRU de la forme : $? \rightarrow (Vente, k\text{€})$ ($s_{seuil} = 0,25$; $c_{seuil} = 0,7$)



1. Introduction
2. Motivations
3. Problème RUDMINE
4. **Algorithme**
✓ **Simulation**
5. Conclusion

Simulation de l'algorithme

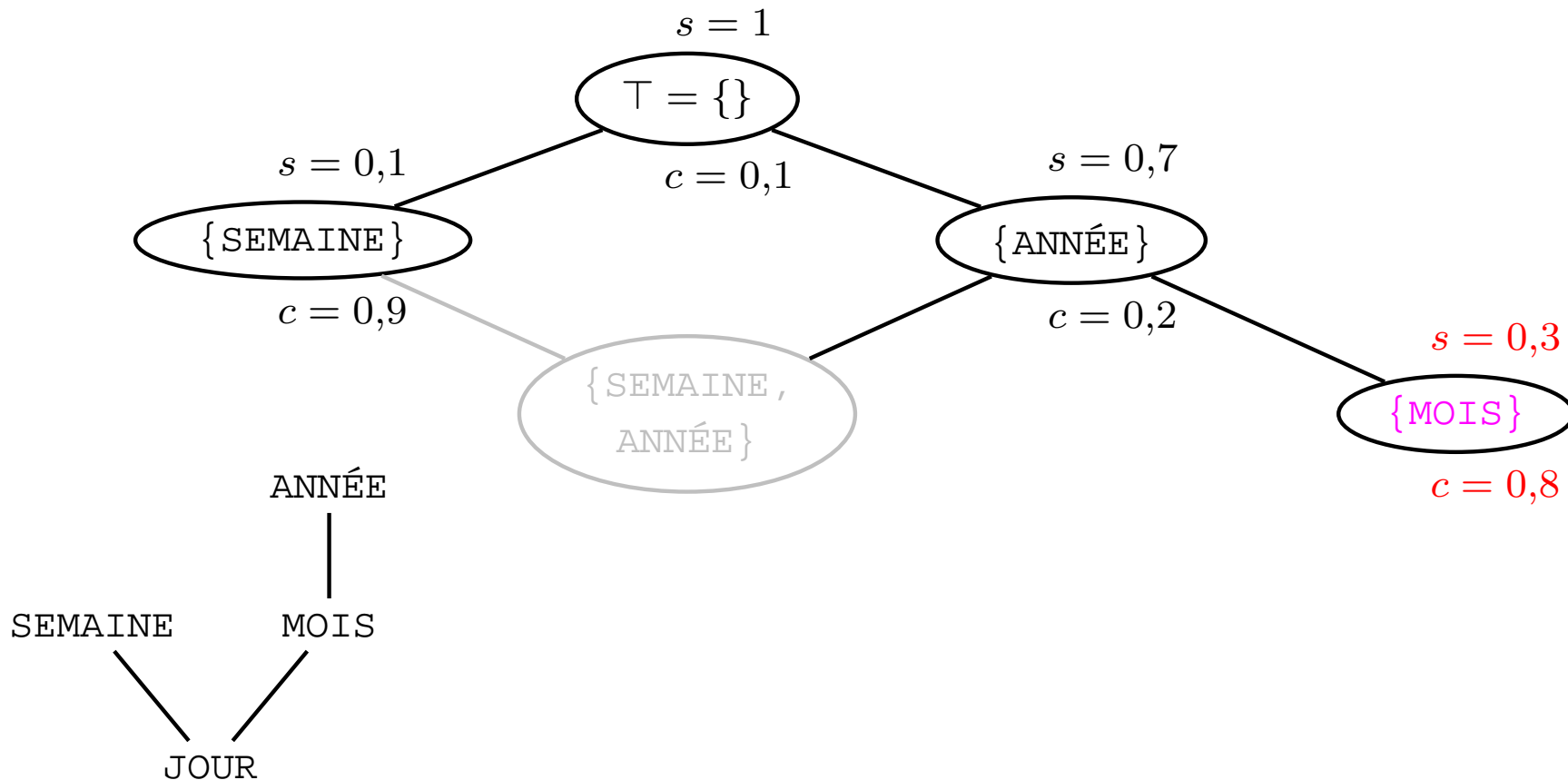
DRU de la forme : $? \rightarrow (Vente, k\text{€})$ ($s_{seuil} = 0,25$; $c_{seuil} = 0,7$)



1. Introduction
2. Motivations
3. Problème RUDMINE
4. **Algorithme**
✓ Simulation
5. Conclusion

Simulation de l'algorithme

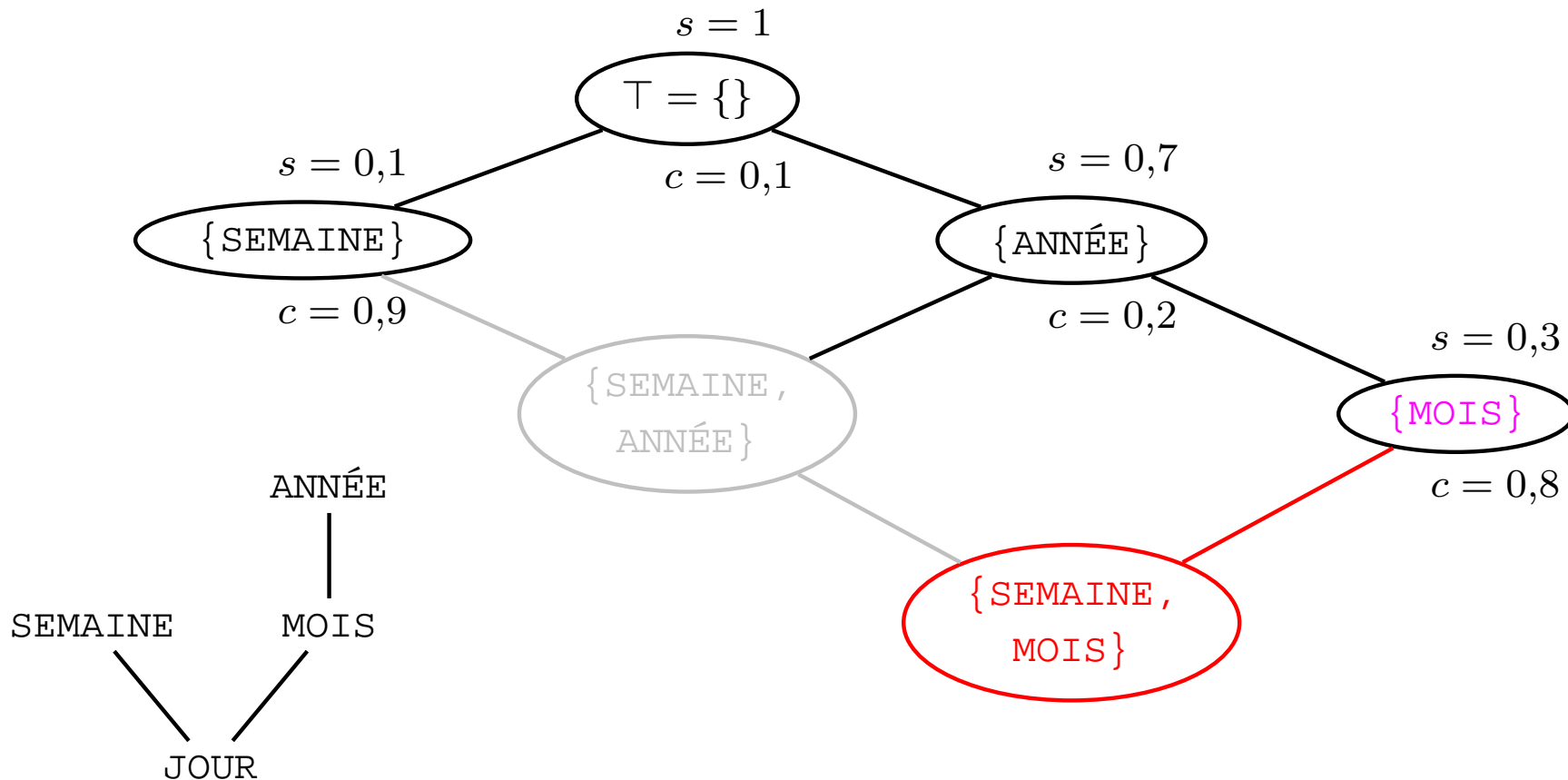
DRU de la forme : $? \rightarrow (Vente, k\text{€})$ ($s_{seuil} = 0,25; c_{seuil} = 0,7$)



1. Introduction
2. Motivations
3. Problème RUDMINE
4. **Algorithme**
✓ Simulation
5. Conclusion

Simulation de l'algorithme

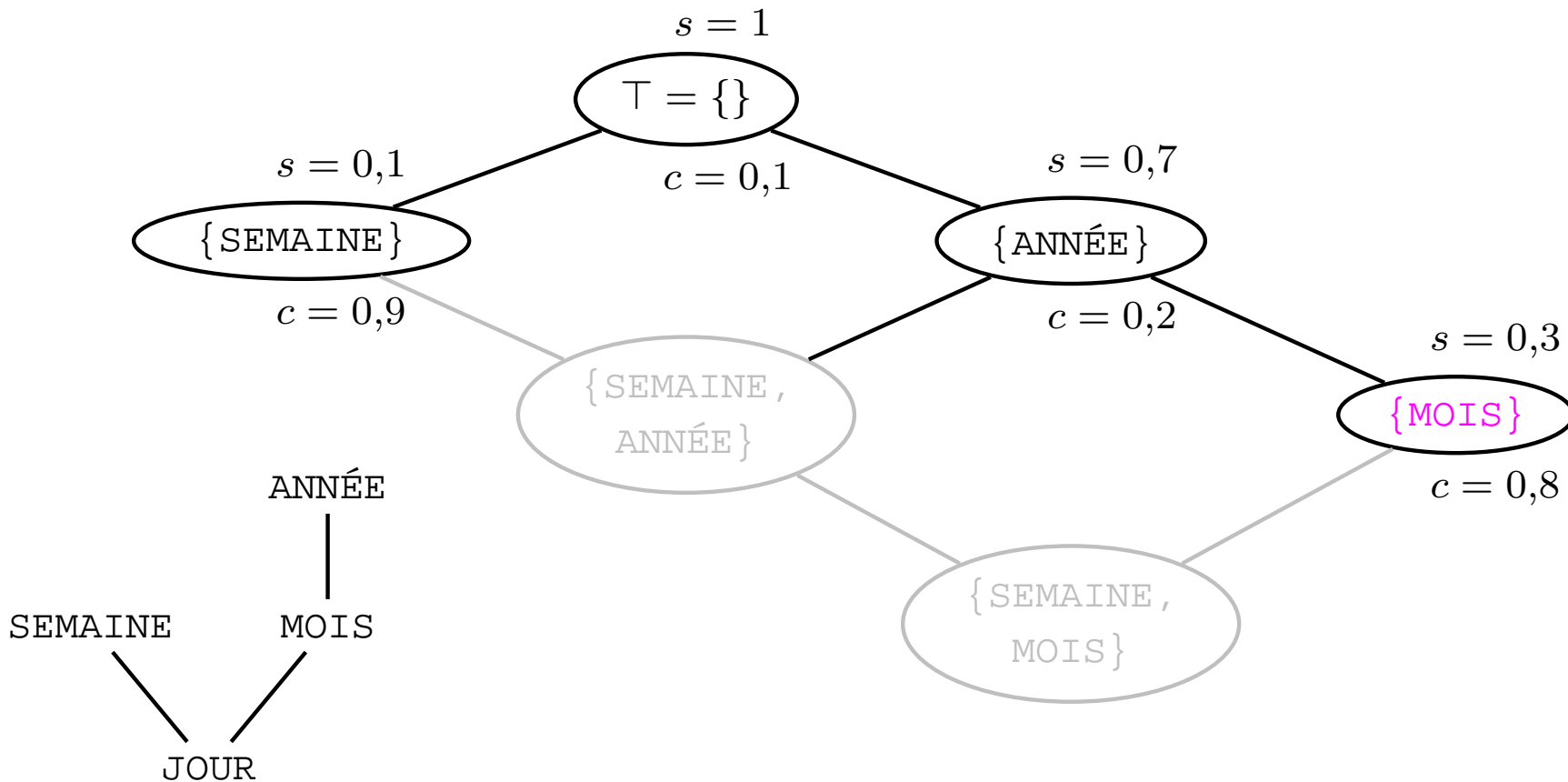
DRU de la forme : $? \rightarrow (Vente, k\text{€})$ ($s_{seuil} = 0,25; c_{seuil} = 0,7$)



1. Introduction
2. Motivations
3. Problème RUDMINE
4. **Algorithme**
✓ Simulation
5. Conclusion

Simulation de l'algorithme

$(Date, MOIS) \rightarrow (Vente, k\text{€})$



1. Introduction
2. Motivations
3. Problème RUDMINE
4. Algorithme
✓ Performances
5. Conclusion

Performances

⌚ Résultats théoriques

- △ NP-complet en le nombre d'attributs
- △ Linéaire en le nombre de n-uplets

⌚ Performances en pratique

- △ Exponentiel en le nombre d'attributs (légère surlinéarité en le nombre de gen-schémas produit)
- △ Linéaire en le nombre de n-uplets

1. Introduction
2. Motivations
3. Problème RUDMINE
4. Algorithme
5. Conclusion
✓ Résumé

En résumé...

- ⑥ Dépendances roll-ups : extension de la notion de dépendance fonctionnelle
- ⑥ Utilisation des hiérarchies de valeurs d'attributs pour l'extraction d'informations
- ⑥ Algorithme efficace

1. Introduction
2. Motivations
3. Problème RUDMINE
4. Algorithme
5. Conclusion
✓ Perspectives

Perspectives

- ⑥ Généralisation de la notion de DRU (distance entre valeurs)
- ⑥ Conception de méthode de mise en forme normale d'un schéma de bases de données (extension de l'algorithme de Boyce-Codd)
- ⑥ Test de l'efficacité des résultats en prédiction