Gestion de données incertaines et de leur provenance

Pierre Senellart

Séminaire *INSERM & Institut TELECOM*

7 octobre 2008

Données incertaines

Nombreuses sources de données incertaines :

- Erreurs de mesure
- Intégration de données de multiples sources
- Processus automatiques imprécis (extraction d'information, traitement du langage naturel...)
- Jugement humain imparfait

Gérer cette imprécision

Objectif

Ne pas faire comme si cette imprécision n'existait pas, et la gérer de façon aussi rigoureuse que possible, tout au long d'un processus (automatique et humain) qui peut être complexe.

En particulier :

- Utiliser des probabilités pour représenter la confiance en les données
- Interroger les données et récupérer des résultats probabiliste
- Permettre d'ajouter, supprimer, modifier des données de manière probabiliste
- Garder tout au long du processus trace de la provenance des données, afin d'assurer la traçabilité

Gérer cette imprécision

Objectif

Ne pas faire comme si cette imprécision n'existait pas, et la gérer de façon aussi rigoureuse que possible, tout au long d'un processus (automatique et humain) qui peut être complexe.

En particulier :

- Utiliser des probabilités pour représenter la confiance en les données
- Interroger les données et récupérer des résultats probabiliste
- Permettre d'ajouter, supprimer, modifier des données de manière probabiliste
- Garder tout au long du processus trace de la provenance des données, afin d'assurer la traçabilité

Plan de l'exposé

Le modèle relationnel

- Données stockées dans des tables
- Chaque table a un schéma précis (type des colonnes)
- Adapté quand l'information est très structurée

Patient	Examen 1	Examen 2	Diagostic
A	23	12	α
B	10	23	β
C	2	4	γ
D	15	15	α
E	15	17	β

Simples annotations probabilistes

Patient	Examen 1	Examen 2	Diagostic	Probabilité
A	23	12	α	0.9
B	10	23	β	0.8
C	2	4	γ	0.2
C	2	14	γ	0.4
D	15	15	α	0.6
D	15	15	β	0.4
E	15	17	β	0.7
E	15	17	α	0.3

- Permet de représenter la confiance dans chaque entrée de la table
- Des algorithmes efficaces pour répondre aux requêtes
- Impossible d'exprimer des dépendances entre entrées

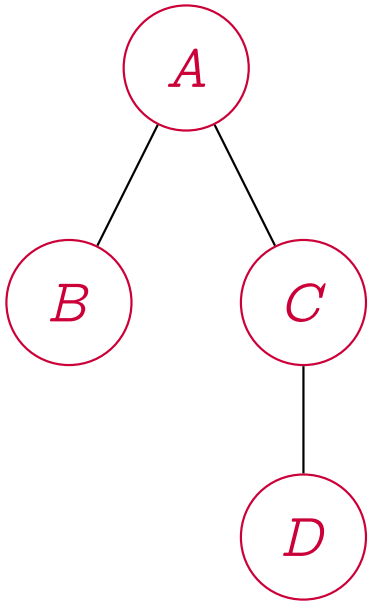
Annotations + contraintes

Patient	Examen 1	Examen 2	Diagostic	Probabilité	
A	23	12	α	0.9	
B	10	23	β	0.8	
C	2	4	γ	0.2	$\oplus$
C	2	14	γ	0.4	
D	15	15	β	0.6	$\oplus$
D	15	15	α	0.4	
E	15	17	β	0.7	$\oplus$
E	15	17	α	0.3	

- Toujours des algorithmes efficaces pour les requêtes
- Dépendances simples (exclusion) exprimables, mais pas dépendances plus complexes

Plan de l'exposé

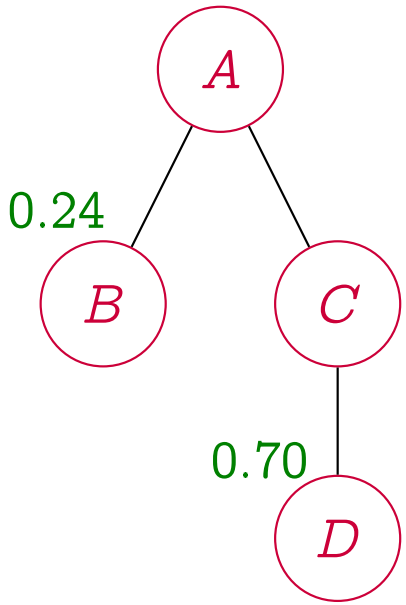
Le modèle semi-structuré




```
<a>
  <b>...</b>
  <c>
    <d>...</d>
  </c>
</a>
```

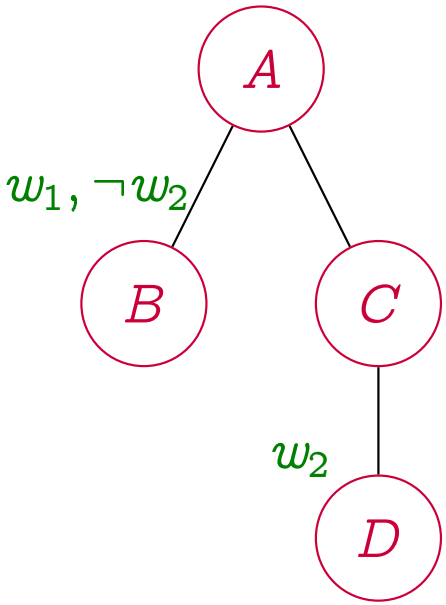
- Présentation arborescente des données
- Pas (ou moins) de contraintes de schéma
- Permet de mêler balises (contenu structuré) et texte (contenu non structuré)
- Particulièrement adapté à du contenu annoté

Simples annotations probabilistes



- Probabilités associées aux nœuds de l'arbre
- Exprime les dépendances entre parent et enfant
- Impossible d'exprimer des dépendances plus complexes
- $\Rightarrow$ tous les ensembles de mondes possibles ne sont pas exprimables de cette façon !

Annotations par variables d'événements



Évén.	Proba.
w_1	0.8
w_2	0.7

sémantique →

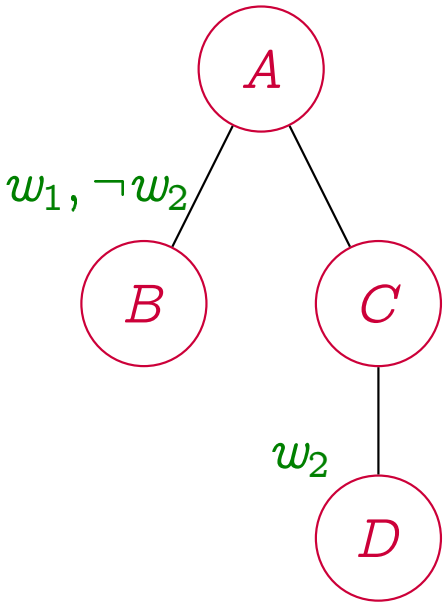
$p_1 = 0.06$

$p_2 = 0.70$

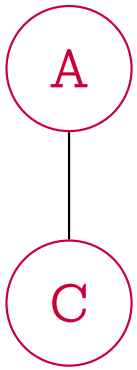
$p_3 = 0.24$

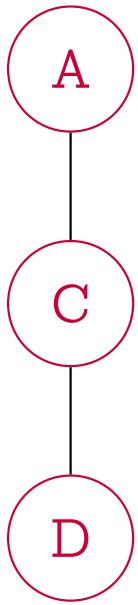
- Expression de dépendances arbitrairement complexes, et algorithmes efficaces pour les requêtes et mises à jours (dans les cas « faciles ») !
- Évidemment, possibilité d'adapter au cas relationnel

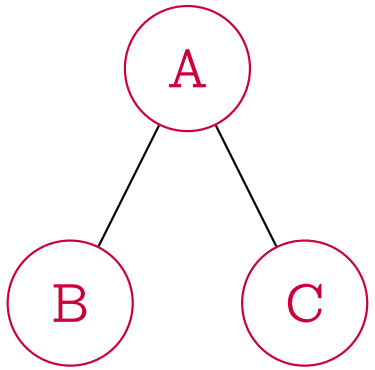
Annotations par variables d'événements



sémantique →







Évén.	Proba.
w_1	0.8
w_2	0.7

$p_1 = 0.06$ $p_2 = 0.70$ $p_3 = 0.24$

- Expression de dépendances arbitrairement complexes, et algorithmes efficaces pour les requêtes et mises à jours (dans les cas « faciles ») !
- Évidemment, possibilité d'adapter au cas relationnel

Gestion de la provenance

- Variables d'événements : peuvent représenter l'origine des données
- Typiquement :
  1. À chaque mise à jour (probabiliste), une nouvelle variable d'évènement est introduite
  2. Ces variables restent présentes tout au long de la vie de la base de données
  3. Les résultats des requêtes sont assorties de probabilités, mais aussi des variables d'évènements correspondantes
- Permet de garder trace, sans coût supplémentaire, de l'origine des données !

Plan de l'exposé

Quelques références

- De nombreux travaux sur les bases de données relationnelles probabilistes

Nilesh Dalvi and Dan Suciu.
Management of Probabilistic Data:
Foundations and Challenges.
Proc. PODS, Beijing, China, June 2007.

- Une synthèse des modèles des bases de données XML probabilistes

Benny Kimelfeld and Yuri Kosharovski and Yehoshua Sagiv.
Query Efficiency in Probabilistic XML Models.
Proc. SIGMOD, Vancouver, Canada, June 2008.

À TELECOM ParisTech

- Intérêt général pour la représentation des données incertaines
- Des travaux passés et en cours sur les modèles probabilistes avec variables d'évènements
- Intérêt particulier pour la gestion des mises à jour
- Thèse à venir sur la gestion de données imprécises dans un système en pair-à-pair auto-administré (projet ANR DataRing)

Pour aller plus loin...

- Gestion des intervalles de valeur, et des distributions continues de probabilité
- Mises à jours complexes (modifications, suppressions) efficaces
- Transformer des valeurs de confiance en vraies probabilités