

Modèle d'incertitude et langage de requêtes pour le DataRing

Pierre Senellart



Réunion de lancement *DataRing*
27 janvier 2009

Nombreuses sources de **données incertaines** :

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

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 probabilistes
- 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

Le tout dans un cadre distribué !

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 **probabilistes**
- 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**

Le tout dans un cadre distribué !

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 **probabilistes**
- 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**

Le tout dans un cadre distribué!

- 1 Données incertaines, processus incertains
- 2 Modèles probabilistes
 - Tables (modèle relationnel)
 - Arbres (modèle semi-structuré)
- 3 Modèle incertain pour le DataRing
- 4 Requêtes

- 1 Données incertaines, processus incertains
- 2 Modèles probabilistes
 - Tables (modèle relationnel)
 - Arbres (modèle semi-structuré)
- 3 Modèle incertain pour le DataRing
- 4 Requêtes

- 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

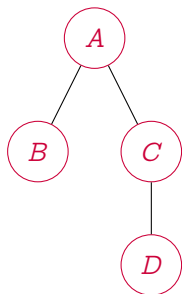
Patient	Examen 1	Examen 2	Diagnostic	Probabilité
A	23	12	α	0.9
B	10	23	β	0.8
C	2	4	γ	0.2
C	2	14	γ	0.4
D	15	15	α	0.6
D	15	15	β	0.4
E	15	17	β	0.7
E	15	17	α	0.3

- 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

Patient	Examen 1	Examen 2	Diagnostic	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	β	0.6
D	15	15	α	0.4
E	15	17	β	0.7
E	15	17	α	0.3

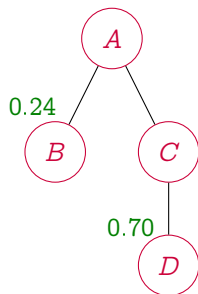
- Toujours des algorithmes **efficaces** pour les requêtes
- **Dépendances** simples (exclusion) exprimables, mais pas dépendances plus complexes

- 1 Données incertaines, processus incertains
- 2 Modèles probabilistes
 - Tables (modèle relationnel)
 - Arbres (modèle semi-structuré)
- 3 Modèle incertain pour le DataRing
- 4 Requêtes



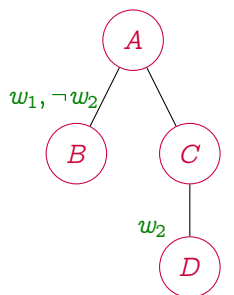
```
<a>  
  <b>...</b>  
  <c>  
    <d>...</d>  
  </c>  
</a>
```

- 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** et **hétérogène**



- 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



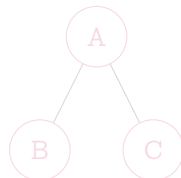
sémantique \longrightarrow



$$p_1 = 0.06$$



$$p_2 = 0.70$$

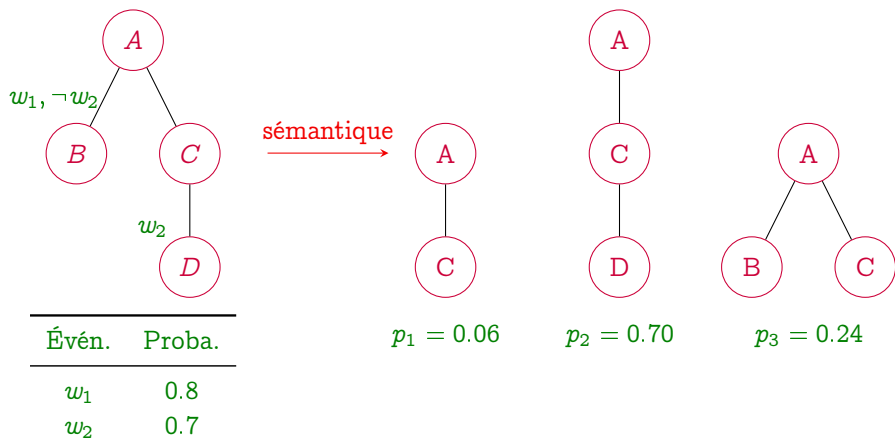


$$p_3 = 0.24$$

Évén.	Proba.
w_1	0.8
w_2	0.7

- 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



- 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

- 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 !

- 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 XML probabilistes



Benny Kimelfeld and Yuri Koscharovski and Yehoshua Sagiv.

Query Efficiency in Probabilistic XML Models.

Proc. SIGMOD, Vancouver, Canada, June 2008.

- 1 Données incertaines, processus incertains
- 2 Modèles probabilistes
- 3 **Modèle incertain pour le DataRing**
- 4 Requêtes

- Pas de schéma centralisé, données **hétérogènes** (\Rightarrow XML ?)
- Modélisation d'information **incomplète** et potentiellement **contradictoire**
- Souvent, **pas de source** explicite de **probabilités**
- Lien entre incertitude au niveau local et au niveau global
- Erreurs de mesure : distribution **continue** de probabilité

- Théorie des **probabilités** : $P(A \cup B) = P(A) + (1 - P(A|B))P(B)$
- Théorie des **possibilités** : $poss(A \cup B) = \max(poss(A), poss(B))$
- Exprime la **possibilité** qu'un événement soit réalisé

Avantage

Dans de nombreux cas, on exprime une **confiance** en l'information, mieux traduite par une **distribution de possibilité** qu'une **distribution de probabilité** (pas besoin de prévoir tous les cas).

Inconvénient

Mathématiquement **moins exploitable** que la théorie des probabilités

- 1 Données incertaines, processus incertains
- 2 Modèles probabilistes
- 3 Modèle incertain pour le DataRing
- 4 Requêtes**

- Au minimum : requêtes à **motif d'arbre** + **jointures** (sous-ensemble classique de XQuery)
- Mais aussi : **interroger les probabilités** elles-mêmes (p. ex., ensemble des résultats avec plus de .1 de probabilité...)
- Requêtes d'**agrégation** ?
- Résumé d'une distribution de probabilité avec l'**espérance** ? la **variance** ?
- Idée générale : **réécriture** des requêtes pour se ramener à de l'évaluation de requêtes XQuery (distribuées).

Attention !

La **projection** rend les requêtes à motif d'arbre #P-complètes (en le nombre de variable d'événements)!

