Probabilistic XML: Survey and Challenges

Pierre Senellart

Webdam Workshop
28 August 2009

Outline

1 Motivation

2 Existing Work

3 Challenges

Uncertain data

Numerous sources of uncertain data:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemata
- Imprecise automatic process (information extraction, natural language processing, etc.)
- Imperfect human judgment

Managing this imprecision

Objective

Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

Especially:

- Use probabilities to represent the confidence in the data
- Query data and retrieve probabilistic results
- Allow adding, deleting, modifying data in a probabilistic way
- (If possible) Keep throughout the process lineage/provenance information, so as to ensure traceability

Managing this imprecision

Objective

Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

Especially:

- Use probabilities to represent the confidence in the data
- Query data and retrieve probabilistic results
- Allow adding, deleting, modifying data in a probabilistic way
- (If possible) Keep throughout the process lineage/provenance information, so as to ensure traceability

Why XML?

- Extensive literature about probabilistic relational databases [DRS09, Wid05, Koc09]
- Different typical querying languages: conjunctive queries vs tree-pattern queries (possibly with joins)
- Cases where a tree-like model might be appropriate:
  - No schema or few constraints on the schema
  - Independent modules annotating freely a content warehouse
  - Inherently tree-like data (e.g., mailing lists) with naturally occurring queries involving the descendant axis

Remark

Some results can be transferred from one model to the other. In other cases, connection much trickier (see later)!

Outline

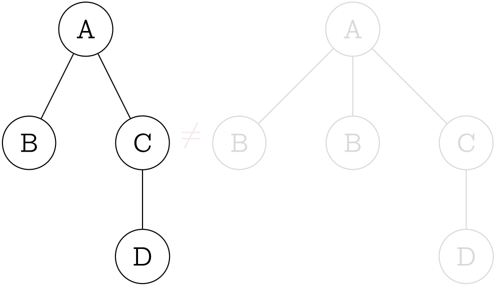
1 Motivation

2 Existing Work

3 Challenges

Trees and possible worlds

Unordered data trees



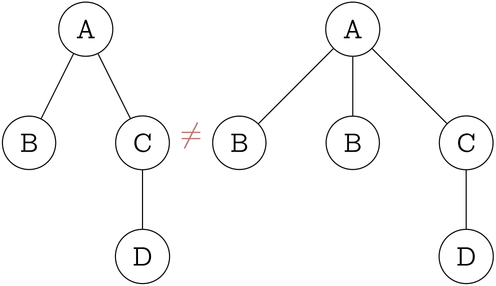
(multiset semantics)

Sample space: Set of all such data trees.

Probabilistic XML database: (Succinct) representation of a discrete probability distribution over this sample space (= a set of possible worlds).

Trees and possible worlds

Unordered data trees



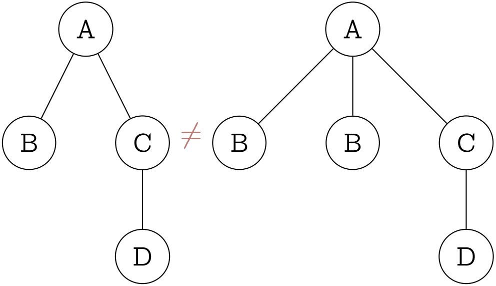
(multiset semantics)

Sample space: Set of all such data trees.

Probabilistic XML database: (Succinct) representation of a discrete probability distribution over this sample space (= a set of possible worlds).

Trees and possible worlds

Unordered data trees

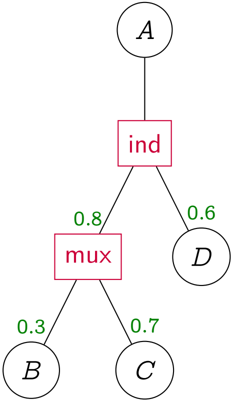


(multiset semantics)

Sample space: Set of all such data trees.

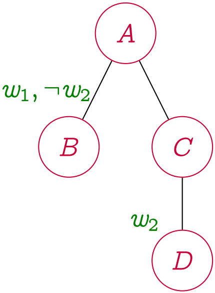
Probabilistic XML database: (Succinct) representation of a discrete probability distribution over this sample space (= a set of possible worlds).

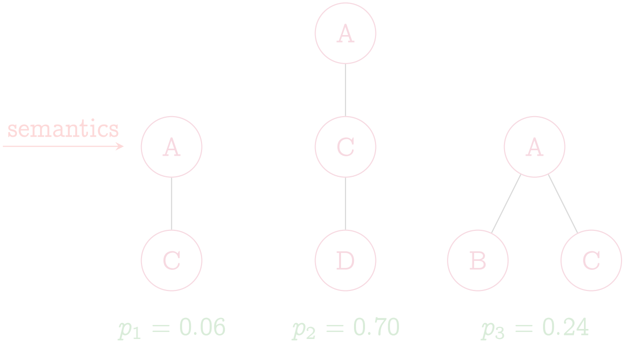
Local dependencies [NJ02]



- Tree with ordinary (circles) and distributional (rectangles) nodes
- Distributional nodes specify how their children can be randomly selected (here, independently or in a mutually exclusive way)
- Possible-world semantics: every possible selection of children of distributional nodes, with associated probability
- No long-distance probabilistic dependencies in the tree!
- Minor generalizations of ind and mux also exist

Arbitrary dependencies [AS06]

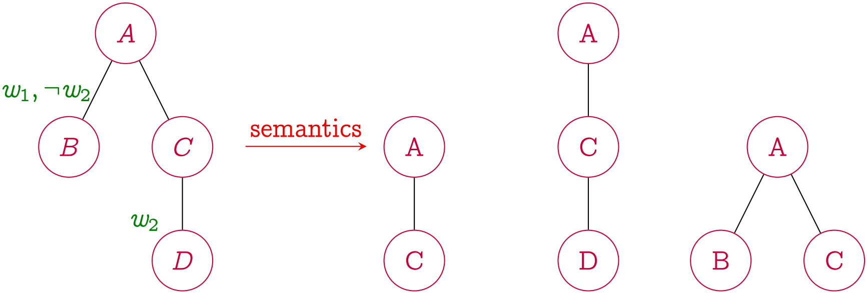




Event	Prob.
w_1	0.8
w_2	0.7

- Conjunctions of independent events on each node of the tree [IL84]
- Expresses arbitrarily complex dependencies
- Bonus: events can track lineage [FGT08]

Arbitrary dependencies [AS06]



Event	Prob.
w_1	0.8
w_2	0.7

- Conjunctions of independent events on each node of the tree [IL84]
- Expresses arbitrarily complex dependencies
- Bonus: events can track lineage [FGT08]

Summary of results (data complexity)

	Local dependencies	Arbitrary dependencies
Expressiveness	Full expressive power [AS06, KKS08]	
Compactness	AD exponentially more compact than LD [Sen07, KKS08]	
Queries		
• tree-pattern	PTIME [KKS09]	$\mathrm{FP}^{\#\mathrm{P}}$-complete [KKS08]
• with joins	$\mathrm{FP}^{\#\mathrm{P}}$-complete	$\mathrm{FP}^{\#\mathrm{P}}$-complete
• project-free	PTIME	PTIME [SA07]
• TP + `HAVING`	PTIME [CKS08]	$\mathrm{FP}^{\#\mathrm{P}}$-complete
Tree automaton (typing, MSO)	PTIME [CKS09]	$\mathrm{FP}^{\#\mathrm{P}}$-complete
Updates	Intractable [AKSS09]	Insertions tractable, Deletions intractable [SA07]

Outline

1 Motivation

2 Existing Work

3 Challenges

Link with probabilistic relational models

Relational case
(Block-independent disjoint model, [DS07])

- Some conjunctive queries are PTIME
- Others are #P-hard
- Complex conditions to separate the two

XML case (Local dependencies)

- Tree pattern queries are PTIME
- Tree pattern queries with (non-trivial) joins are #P-hard

- Why does the XML case seem simpler?
- Is there some insight to be gained from one case to the other?
- Translating XML data and queries to the relational case yields queries with self-joins, a less well-understood setting

Link with probabilistic relational models

Relational case
(Block-independent disjoint model, [DS07])

- Some conjunctive queries are PTIME
- Others are #P-hard
- Complex conditions to separate the two

XML case (Local dependencies)

- Tree pattern queries are PTIME
- Tree pattern queries with (non-trivial) joins are #P-hard

- Why does the XML case seem simpler?
- Is there some insight to be gained from one case to the other?
- Translating XML data and queries to the relational case yields queries with self-joins, a less well-understood setting

Continuous probability distributions

- Most probabilistic database models assume discrete probabilistic distributions
- Sensor networks, unknown values: need for continuous distributions! (uniform, Gaussian, Poisson, etc.)
- Some existing works on query answering over continuous distributions [CKP03, DGM+04] but no clear semantics
- Claim: this is not more difficult than the discrete case, as long as integration/differentiation are easy (symbolically or numerically) for the considered distributions
- Discrete distributions can be modeled as Diracs

Tractable extensions of the local dependency model

- Arbitrary dependencies: not tractable
- Local dependencies: not practical
- Somewhere in between?
  - What makes the arbitrary dependency model hard?
  - How can the local dependency model be generalized, while remaining tractable?
- And can we go further? cf. XML schemas
  - Trees of unbounded depth
  - Trees of unbounded width
  - Infinite trees?

But where do probabilities come from?!

- Do the numbers assigned as probabilities in PDBMS really make sense?
- In some cases, sources of “good” probabilities:
  - Statistics
  - Conditional Random Fields
- What about the rest? Does it really make sense to model uncertainty with probabilities?

A system that just works

- Nothing else than toy systems exist for probabilistic XML
- What should it be based upon:
  - a probabilistic relational DBMS?
  - a native XML DBMS?
- Systems issue: distribution, indexing, etc.
- And need for a killer application! Probabilistic content warehouse?

Merci.

References I

Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, and Pierre Senellart.
On the expressiveness of probabilistic XML models.
VLDB Journal, 2009.
To appear.

Serge Abiteboul and Pierre Senellart.
Querying and updating probabilistic information in XML.
In *Proc. EDBT*, Munich, Germany, March 2006.

Reynold Cheng, Dmitri V. Kalashnikov, and Sunil Prabhakar.
Evaluating probabilistic queries over imprecise data.
In *Proc. SIGMOD*, San Diego, CA, USA, June 2003.

References II

Sara Cohen, Benny Kimelfeld, and Yehoshua Sagiv.
Incorporating constraints in probabilstic XML.
In *Proc. PODS*, Vancouver, BC, Canada, June 2008.

Sara Cohen, Benny Kimelfeld, and Yehoshua Sagiv.
Running tree automata on probabilistic XML.
In *Proc. PODS*, Providence, RI, USA, June 2009.

Amol Deshpande, Carlos Guestrin, Samuel Madden, Joseph M. Hellerstein, and Wei Hong.
Model-driven data acquisition in sensor networks.
In *Proc. VLDB*, Toronto, ON, Canada, August 2004.

Nilesh Dalvi, Chrisopher Ré, and Dan Suciu.
Probabilistic databases: Diamonds in the dirt.
Communications of the ACM, 52(7), 2009.

References III

Nilesh N. Dalvi and Dan Suciu.
Management of probabilistic data: foundations and challenges.
In *Proc. PODS*, Beijing, China, June 2007.

J. Nathan Foster, Todd J. Green, and Val Tannen.
Annotated XML: queries and provenance.
In *Proc. PODS*, Vancouver, BC, Canada, June 2008.

Tomasz Imieliński and Witold Lipski.
Incomplete information in relational databases.
Journal of the ACM, 31(4):761–791, 1984.

Benny Kimelfeld, Yuri Kosharovsky, and Yehoshua Sagiv.
Query efficiency in probabilistic XML models.
In *Proc. SIGMOD*, Vancouver, BC, Canada, June 2008.

References IV

Benny Kimelfeld, Yuri Kosharovsky, and Yehoshua Sagiv.
Query evaluation over probabilistic XML.
The VLDB Journal, 2009.
To appear.

Christoph Koch.
MayBMS: A system for managing large uncertain and probabilistic databases.
In Charu Aggarwal, editor, *Managing and Mining Uncertain Data*. Springer-Verlag, 2009.

Andrew Nierman and H. V. Jagadish.
ProTDB: Probabilistic data in XML.
In *Proc. VLDB*, Hong Kong, China, August 2002.

References V

Pierre Senellart and Serge Abiteboul.
On the complexity of managing probabilistic XML data.
In *Proc. PODS*, Beijing, China, June 2007.

Pierre Senellart.
Comprendre le Web caché. Understanding the Hidden Web.
PhD thesis, Université Paris-Sud 11, December 2007.

Jennifer Widom.
Trio: A system for integrated management of data, accuracy, and lineage.
In *Proc. CIDR*, Asilomar, CA, USA, January 2005.