On the Expressiveness of Probabilistic XML Models

Serge Abiteboul[1] Benny Kimelfeld[2]
Yehoshua Sagiv[3] Pierre Senellart[4]

1 INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE INRIA centre de recherche SACLAY – ÎLE-DE-FRANCE

2 IBM

4 TELECOM ParisTech

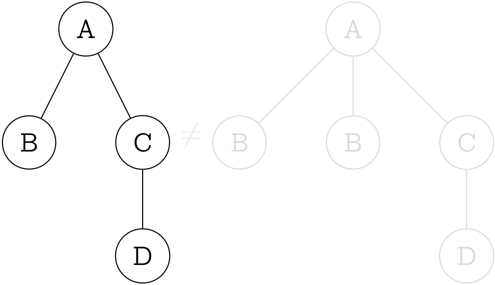
3 האוניברסיטה העברית בירושלים The Hebrew University of Jerusalem

Dagstuhl Seminar on *Uncertainty Management*, 14 October 2008

Probabilistic XML

Framework

- Unordered data trees
- Details: no attributes, no mixed content. . .



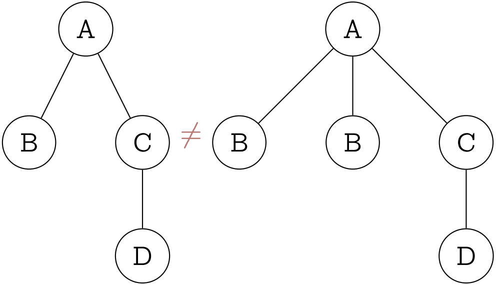
(multiset semantics)

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).

Probabilistic XML

Framework
- Unordered data trees
- Details: no attributes, no mixed content…



(multiset semantics)

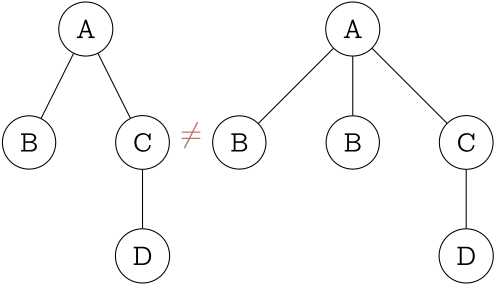
Sample space: Set of all such data trees.

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

Probabilistic XML

Framework

- Unordered data trees
- Details: no attributes, no mixed content...



(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).

A Unifying Framework for Probabilistic XML Models

Goal

- A generic framework for probabilistic XML
- Previously proposed models: concrete instances of this framework
- Comparison of the expressiveness of various models
- Update capabilities in various models
- Efficiency issues

Outline

1 Introduction

2 P-documents

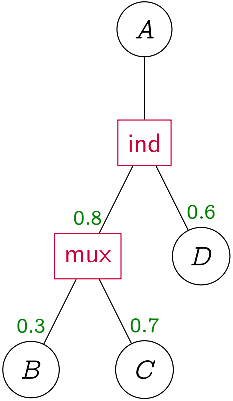
- The P-document Model
- Types of Distributional Nodes
- Link with Previously Studied Models

3 Efficient Translations between Models

4 Update Capabilities

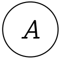
5 Conclusion

P-Documents

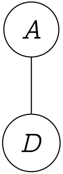


- Tree with ordinary (circles) and distributional (rectangles) nodes
- Distributional nodes specify how their children can be randomly selected
- Several kinds of distributional nodes (see later on)
- Possible-world semantics: every possible selection of children of distributional nodes, with associated probability

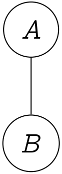
Possible-world semantics



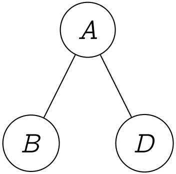
$p_1 = 0.08$



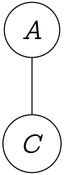
$p_2 = 0.12$



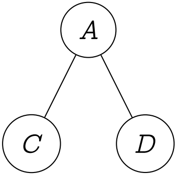
$p_3 = 0.096$



$p_4 = 0.144$

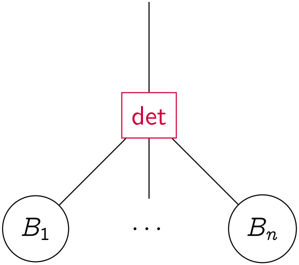


$p_5 = 0.224$



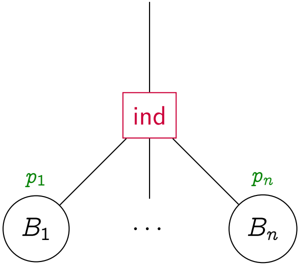
$p_6 = 0.336$

Deterministic



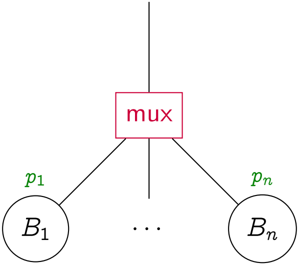
- All children are always chosen
- Not really a distributional node, but sometimes useful in hierarchies

Independent



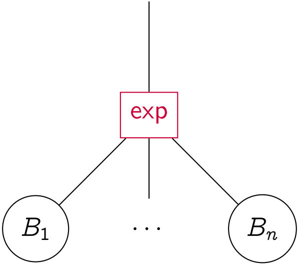
- Children are randomly chosen independently of one another
- The probability of choosing each child is given

Mutually Exclusive



- At most one child is randomly chosen, with probability p_i
- $\sum_{i=1}^{n} p_i \leq 1$

Exponential

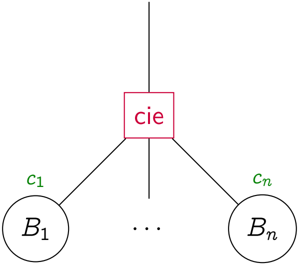


Subset	Probability
W_1	p_1
$\dots$	
W_k	p_k

- The probability of choosing a given subset W_j of the set of children is given
- $1 \leq k \leq 2^n$
- $\sum_{j=1}^{k} p_j = 1$

Conjunction of Independent Events

Event	Probability
e_1	p_1
	$\dots$
e_k	p_k



- Each c_i is a conjunction of the random events e_j or their negations $\neg e_j$ (e.g., $e_1 \wedge \neg e_2 \wedge e_3$)
- Each e_j is independent of the other ones
- The probability of each e_j is given
- The e_j can be shared across multiple distributional nodes
- Only distributional node expressing long-distance dependency
- Reminiscent of [Imieliński and Lipski, 1984]

Families of P-documents

Definition

- $\mathsf{PrXML}^{\{\mathrm{type}_1,\mathrm{type}_2,\ldots\}}$: family of p-documents obtained with the distributional nodes $\mathrm{type}_1, \mathrm{type}_2, \ldots$
- $\mathsf{PrXML}^{\{\mathrm{type}_1,\mathrm{type}_2,\ldots\}}_{|\not h}$: no hierarchy of distributional nodes is allowed (i.e., the child of a distributional node is an ordinary node)

Link with Previously Studied Models

[Nierman and Jagadish, 2002]: $\mathsf{PrXML}^{\{\mathsf{mux},\mathsf{det}\}}$

[van Keulen et al., 2005]: subset of $\mathsf{PrXML}^{\{\mathsf{mux},\mathsf{det}\}}$

[Abiteboul and Senellart, 2006, Senellart and Abiteboul, 2007]: $\mathsf{PrXML}^{\{\mathsf{ind}\}}_{|\not{h}}$ and $\mathsf{PrXML}^{\{\mathsf{cie}\}}_{|\not{h}}$

[Hung et al., 2003]: subset of $\mathsf{PrXML}^{\{\mathsf{exp}\}}_{|\not{h}}$ (graphs restricted to trees)

[Hung et al., 2007]: subset of $\mathsf{PrXML}^{\{\mathsf{exp}\}}_{|\not{h}}$ (intervals restricted to points)

Outline

1 Introduction

2 P-documents

3 Efficient Translations between Models
- V-translations
- Main Results

4 Update Capabilities

5 Conclusion

V-translations

Definition

$\mathcal{F}$ is v-translatable to $\mathcal{F}'$ if, for each $\tilde{\mathcal{P}} \in \mathcal{F}$, there is a $\tilde{\mathcal{P}}' \in \mathcal{F}'$ such that the possible-world semantics of $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{P}}'$ are isomorph.
If, additionally, $\tilde{\mathcal{P}}'$ can be obtained from $\tilde{\mathcal{P}}$ in polynomial time, $\mathcal{F}$ is efficiently v-translatable to $\mathcal{F}'$.

Main Results

- All families of p-documents are translatable to $\mathsf{PrXML}^{\{\mathsf{mux,det}\}}$
- $\mathsf{PrXML}^{\{\mathsf{mux,ind}\}}$ is efficiently translatable to $\mathsf{PrXML}^{\{\mathsf{exp}\}}$ and to $\mathsf{PrXML}^{\{\mathsf{cie}\}}$
- $\mathsf{PrXML}^{\{\mathsf{exp}\}}$ is not efficiently translatable to $\mathsf{PrXML}^{\{\mathsf{exp}\}}_{|\not h}$
- $\mathsf{PrXML}^{\{\mathsf{cie}\}}$ is efficiently translatable to $\mathsf{PrXML}^{\{\mathsf{cie}\}}_{|\not h}$
- $\mathsf{PrXML}^{\{\mathsf{cie}\}}$ is not efficiently translatable to $\mathsf{PrXML}^{\{\mathsf{ind,mux,exp}\}}$
- $\mathsf{PrXML}^{\{\mathsf{exp}\}}$ to $\mathsf{PrXML}^{\{\mathsf{cie}\}}$: open problem, but $\mathsf{PrXML}^{\{\mathsf{exp}\}}$ with bounded height or bounded degree efficiently translatable to $\mathsf{PrXML}^{\{\mathsf{cie}\}}$

Outline

1 Introduction

2 P-documents

3 Efficient Translations between Models

4 Update Capabilities
- V-updates
- Main Results

5 Conclusion

V-updates

- Elementary insertions and deletions
- Locator query indicating where to apply the (cf. XPath for XUpdate, XQuery for XQuery Update)
- The update itself can be probabilistic: “Insert this subtree at each node matched by this query with probability p.”

Main Results

- $\mathsf{PrXML}^{\{\mathsf{mux},\mathsf{det}\}}$ (and any family v-translatable to this) closed under updates for any class of queries
- $\mathsf{PrXML}^{\{\mathsf{cie}\}}$, $\mathsf{PrXML}^{\{\mathsf{mux}\}}$, $\mathsf{PrXML}^{\{\mathsf{exp}\}}$, etc.: tractably closed under updates defined by single-path queries such that the matched node is at the end of the path
- Without cie nodes: not tractably closed under insertions defined by single-path queries
- $\mathsf{PrXML}^{\{\mathsf{cie}\}}$: tractably closed under insertions defined by tree-pattern queries with joins

Outline

1 Introduction

2 P-documents

3 Efficient Translations between Models

4 Update Capabilities

5 Conclusion

A Word About Queries

It was shown [Kimelfeld and Sagiv, 2007, Kimelfeld et al., 2008] that:

- Tree-pattern projection queries can be processed efficiently in $\mathsf{PrXML}^{\{\mathsf{ind},\mathsf{mux}\}}$.
- Tree-pattern projection queries are #P-complete in $\mathsf{PrXML}^{\{\mathsf{cie}\}}$.

In summary:

- $\mathsf{PrXML}^{\{\mathsf{cie}\}}$ is more succinct.
- Simple updates remain tractable in $\mathsf{PrXML}^{\{\mathsf{cie}\}}$.
- . . . but (projection) queries are intractable in $\mathsf{PrXML}^{\{\mathsf{cie}\}}$.

Trade-off between queries and updates, or between queries and expressibility of complex dependencies.

A Word About Queries

It was shown [Kimelfeld and Sagiv, 2007, Kimelfeld et al., 2008] that:

- Tree-pattern projection queries can be processed efficiently in $\mathsf{PrXML}^{\{\mathsf{ind},\mathsf{mux}\}}$.
- Tree-pattern projection queries are #P-complete in $\mathsf{PrXML}^{\{\mathsf{cie}\}}$.

In summary:

- $\mathsf{PrXML}^{\{\mathsf{cie}\}}$ is more succinct.
- Simple updates remain tractable in $\mathsf{PrXML}^{\{\mathsf{cie}\}}$.
- . . . but (projection) queries are intractable in $\mathsf{PrXML}^{\{\mathsf{cie}\}}$.

Trade-off between queries and updates, or between queries and expressibility of complex dependencies.

A Word About Queries

It was shown [Kimelfeld and Sagiv, 2007, Kimelfeld et al., 2008] that:

- Tree-pattern projection queries can be processed efficiently in $\mathsf{PrXML}^{\{\mathsf{ind},\mathsf{mux}\}}$.
- Tree-pattern projection queries are #P-complete in $\mathsf{PrXML}^{\{\mathsf{cie}\}}$.

In summary:

- $\mathsf{PrXML}^{\{\mathsf{cie}\}}$ is more succinct.
- Simple updates remain tractable in $\mathsf{PrXML}^{\{\mathsf{cie}\}}$.
- … but (projection) queries are intractable in $\mathsf{PrXML}^{\{\mathsf{cie}\}}$.

Trade-off between queries and updates, or between queries and expressibility of complex dependencies.

Perspectives

- Alternative approach: external constraints [Cohen et al., 2008]
- Multiset → set semantics
- Equivalence of p-documents
- Validation against a DTD

Merci.

References I

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

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

Edward Hung, Lise Getoor, and V. S. Subrahmanian. PXML: A probabilistic semistructured data model and algebra. In *Proc. ICDE*, Bangalore, India, March 2003.

Edward Hung, Lise Getoor, and V. S. Subrahmanian. Probabilistic interval XML. *ACM Transactions on Computational Logic*, 8(4), 2007.

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

References II

Benny Kimelfeld and Yehoshua Sagiv. Matching twigs in probabilistic XML. In *Proc. VLDB*, Vienna, Austria, September 2007.

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

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

Pierre Senellart and Serge Abiteboul. On the complexity of managing probabilistic XML data. In *Proc. PODS*, pages 283–292, Beijing, China, June 2007.

Maurice van Keulen, Ander de Keijzer, and Wouter Alink. A probabilistic XML approach to data integration. In *Proc. ICDE*, April 2005.