

Uncertain Version Control in Open Collaborative Editing of Tree-Structured Documents

M. Lamine BA, Talel Abdessalem & Pierre Senellart

<http://dbweb.enst.fr/>

13th ACM Symposium on DocEng – Sept 10-13, Florence (Italy)





Version Control of Uncertain Data

- **Data in large-scale, open and collaborative editing platforms, such as Wikipedia, are inherently uncertain**
 - contributors with different reliability, conflicts, malicious edits, . . .
 - need version control to maintain the quality of document versions
- Existing version control approaches are all deterministic
 - no room to uncertainty handling in the versioning process

👉 A version control model aware of uncertain data may be helpful

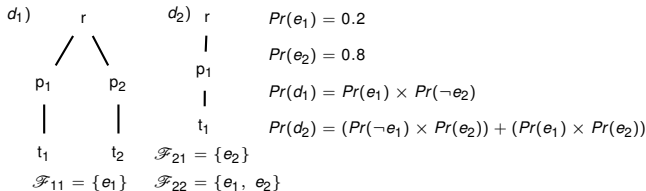
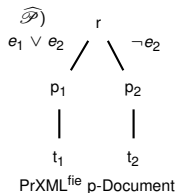




Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]

- Unordered, unranked, and labeled **XML trees** with annotated edges
 - annotations are **propositional formulas** of random Boolean variables



Possible worlds and their probabilities

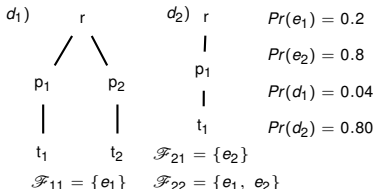
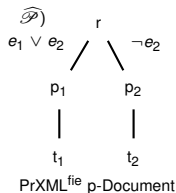




Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]

- Unordered, unranked, and labeled **XML trees** with annotated edges
 - annotations are **propositional formulas** of random Boolean variables



Possible worlds and their probabilities

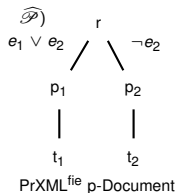




Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]

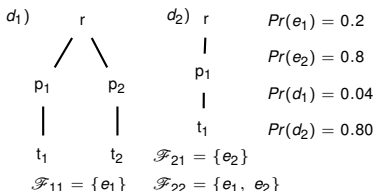
- Unordered, unranked, and labeled **XML trees** with annotated edges
 - annotations are **propositional formulas** of random Boolean variables



- Enumerating all possible worlds and their probabilities

- Enable also to model uncertain updates on (uncertain) nodes [Kharlamov et al.(2010)]

- ☞ Integrate such a representation in a typical version control process



Possible worlds and their probabilities



Uncertain Multi-Version XML Document

Uncertain Version Control Model

Semantics of Updates

Evaluation of the model

Performance Analysis

Filtering capabilities





Uncertain Multi-Version XML Document

Uncertain Version Control Model

Defines two equivalent views over any uncertain multi-version XML tree

- set \mathcal{V} of random variables $e_0, e_1 \dots e_n$ modeling the tree states
- infinite set \mathcal{D} of all (unordered) XML trees including the versions

$\mathcal{G}(\mathcal{G}, \Omega)$: Logical View

$\widehat{\mathcal{P}}(\mathcal{G}, \widehat{\mathcal{P}})$: Probabilistic XML Encoding

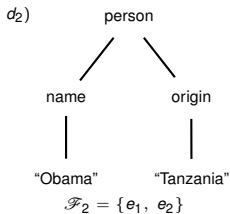
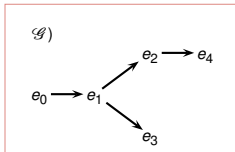
- DAG \mathcal{G} built on variables in \mathcal{V}
- Mapping $\Omega : 2^{\mathcal{V} \setminus \{e_0\}} \rightarrow \mathcal{D}$ which computes the possible versions according to sets of valid events
- Similar DAG \mathcal{G} of random variables in \mathcal{V}
- Probabilistic XML tree $\widehat{\mathcal{P}}$ which defines the same probability distribution as Ω mapping



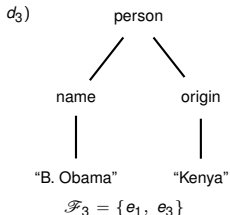
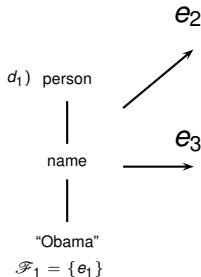
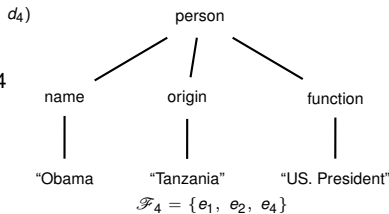


Uncertain Multi-Version XML Document

Uncertain Version Control Model (Example)



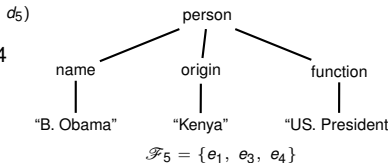
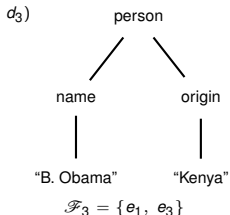
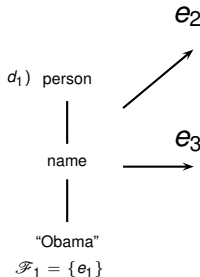
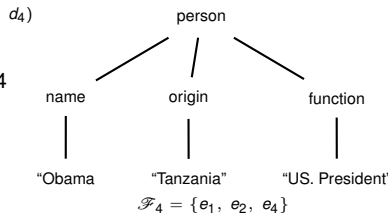
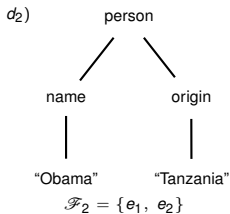
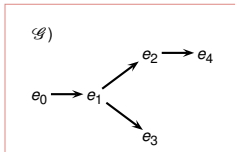
e_4





Uncertain Multi-Version XML Document

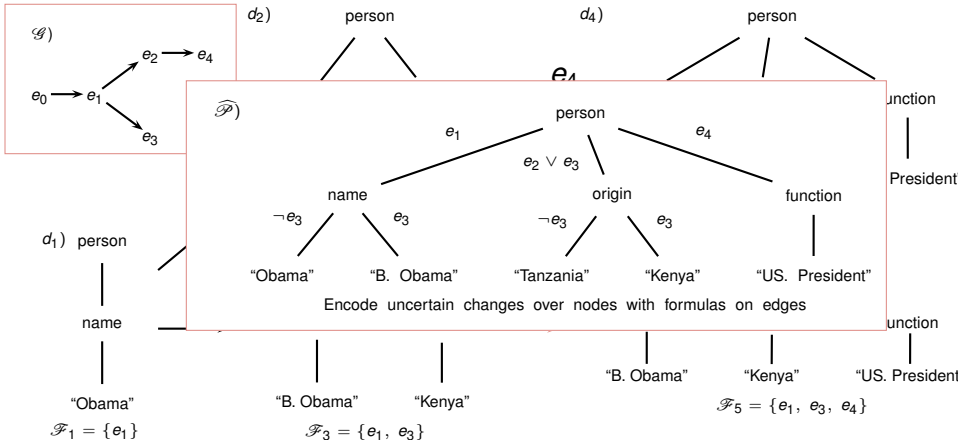
Uncertain Version Control Model (Example)





Uncertain Multi-Version XML Document

Uncertain Version Control Model (Example)





Uncertain Multi-Version XML Document

Semantics of Updates

- Assume an event e_i in \mathcal{G} pointing to the edited version
- Come with a new event e_j *not in* \mathcal{G} and an edit script Δ

Logical definition

Input: (\mathcal{G}, Ω) , e_i , e_j , Δ

- $\mathcal{G} := \mathcal{G} \cup (\{e_j\}, \{(e_i, e_j)\})$
- Extension of Ω to a Ω' mapping



For each event set $\mathcal{F} \in 2^{(\mathcal{V} \setminus \{e_0\}) \cup \{e_j\}}$:

- $\Omega'(\mathcal{F}) = [\Omega(\mathcal{F} \setminus \{e_j\})]^\Delta$ if $e_j \in \mathcal{F}$
- $\Omega'(\mathcal{F}) = [\Omega(\mathcal{F})]$ if $e_j \notin \mathcal{F}$

Probabilistic XML Update

Input: $(\mathcal{G}, \widehat{\mathcal{P}})$, e_i , e_j , Δ

- $\mathcal{G} := \mathcal{G} \cup (\{e_j\}, \{(e_i, e_j)\})$
- Updating $\widehat{\mathcal{P}}$ with operations in Δ

For an insert of x and a delete of y :

- $fie(x) := fie(x) \vee (e_j)$ if $x \in \widehat{\mathcal{P}}$ or insert x in $\widehat{\mathcal{P}}$ with $fie(x) := (e_j)$
- $fie(y) := fie(y) \wedge \neg(e_j)$



Uncertain Multi-Version XML Document

Uncertain Version Control Model

Semantics of Updates

Evaluation of the model

Performance Analysis

Filtering capabilities





Evaluation of the model

Performance Analysis

Estimation of two main metrics: commit time and checkout cost

Baseline Systems

- Versioning tools SubVersion and Git
 - Use of their Java implementations based on the APIs SvnKit and JGit

Real Datasets

History of commits over two large file systems (shared tree-structured data)

- Linux kernel development
- Cassandra project

Set up our system (PrXML) in Java language

Measures are obtained with all accesses in RAM Disk





Evaluation of the model

Performance Analysis (Results)

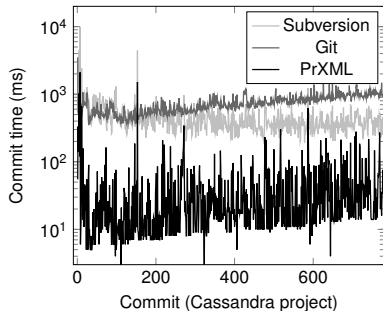
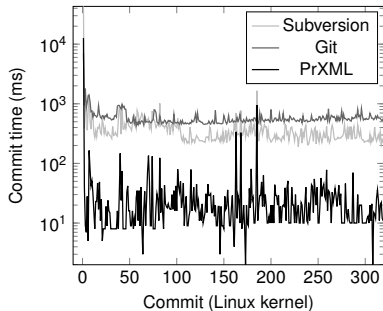


Figure : commit time over real-world datasets (logarithmic y-axis)





Evaluation of the model

Performance Analysis (Results)

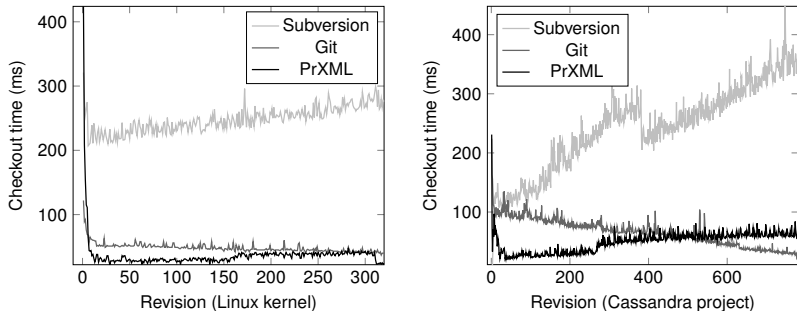


Figure : checkout time over real-world datasets (linear axes)





Evaluation of the model

Filtering capabilities

- ▶ Tests are run over a sample of articles from the Wikipedia dump
 - Automatic filtering of unreliable content, e.g. spams, in versions of articles
 - Generate arbitrary versions that fit user preference
 - ▶ versions from trustworthy authors
 - Test more advanced operations over critical articles such as vandalized pages
 - ▶ e.g. study the impact of considering as reliable some versions affected by vandalism in the history of the edition of a given article
 - Detection of vandalism as well as Wikipedia robots do, automatically manage it while keeping all uncertain versions available for checkout.
 - etc.





Evaluation of the model

Filtering capabilities (Demo [Ba et al.(2011)])

A keyword-based search engine for Wikipedia pages

Keywords: Cedric Villani

Sample articles: Sample articles Online articles

Searching options:

Filtering options on revisions

Extracting the revisions of a given page

Successive revisions	Authors of revisions	Date of edition
Cedric Villani18	1112.128.173	2010-00-01
Cedric Villani17	101.183.23.101	2010-00-01
Cedric Villani16	14pdr	2010-00-08
Cedric Villani15	14pdr	2010-00-08
Cedric Villani14	Automator	2010-00-08
Cedric Villani13	Louismor17	2010-00-01
Cedric Villani12	Louismor17	2010-00-01
Cedric Villani11	1206.63	2010-00-01
Cedric Villani10	82.204.165.103	2010-00-08
Cedric Villani9	Louismor17	2010-00-01
Cedric Villani8	Louismor17	2010-00-01
Cedric Villani7	Louismor17	2010-00-01
Cedric Villani6	Louismor17	2010-00-01
Cedric Villani5	Louismor17	2010-00-01
Cedric Villani4	Louismor17	2010-00-01
Cedric Villani3	Louismor17	2010-00-01
Cedric Villani2	Louismor17	2010-00-01
Cedric Villani1	Montaigne	2011-00-06

Defining reputation values

Selecting list of revisions to merge

(a) keyword-based search engine

Keywords: Cedric Villani

Sample articles: Sample articles Online articles

Range of Revisions:

Building one's own Wikipedia article

Selecting list of revisions to merge

Successive revisions	Authors of revisions	Date of edition
Cedric Villani18	1112.128.173	2010-00-01
Cedric Villani17	101.183.23.101	2010-00-01
Cedric Villani16	14pdr	2010-00-08
Cedric Villani15	14pdr	2010-00-08
Cedric Villani14	Automator	2010-00-08
Cedric Villani13	Louismor17	2010-00-01
Cedric Villani12	Louismor17	2010-00-01
Cedric Villani11	1206.63	2010-00-01
Cedric Villani10	82.204.165.103	2010-00-08
Cedric Villani9	Louismor17	2010-00-01
Cedric Villani8	Louismor17	2010-00-01
Cedric Villani7	Louismor17	2010-00-01
Cedric Villani6	Louismor17	2010-00-01
Cedric Villani5	Louismor17	2010-00-01
Cedric Villani4	Louismor17	2010-00-01
Cedric Villani3	Louismor17	2010-00-01
Cedric Villani2	Louismor17	2010-00-01
Cedric Villani1	Montaigne	2011-00-06

(b) generation of arbitrary versions

Operations on revisions

- Define an available author
- Display merged revision
- Define merge method
- Add further revisions to merge

Operations on reputation values

Authors: Louismor17, 1206.63, 14pdr, 101.183.23.101, 1112.128.173, 82.204.165.103, Automator, Montaigne

Operations on filtering options

Filtering by authors: Louismor17, 1206.63, 14pdr, 101.183.23.101, 1112.128.173, 82.204.165.103, Automator, Montaigne

Filtering by reputation values: Reputation, User-defined

Filtering by user-defined: User-defined

Operations on filtering options

Filtering by authors: Louismor17, 1206.63, 14pdr, 101.183.23.101, 1112.128.173, 82.204.165.103, Automator, Montaigne

Filtering by reputation values: Reputation, User-defined

Filtering by user-defined: User-defined

CEDRIC VILLANI

Contenu (19) (1112.128.173) Cedric Villani (dans 5 October 0010) is a French mathematician working primarily on partial differential equations and mathematical physics. He was awarded the Fields Medal in 2010.

Contenu (19) (1112.128.173) Cedric Villani (dans 5 October 0010) is a French mathematician working primarily on partial differential equations and mathematical physics. He was awarded the Fields Medal in 2010.

Contents

- Work
- References

Work

Villani has worked on the theory of partial differential equations involved in optimal transport, especially the Monge-Kantorovich problem, where, with Leonor Sanz-Solera, he won the first prize for best contemporary research for initial values not more regularized. He has also worked with George Borot on new results with Leonor Sanz-Solera. He has also worked on nonlinear Schrödinger equation. He has worked on the theory of optimal transport and its application to mathematical physics, and with Jean Dolbeault he obtained a series of beautiful Sobolev inequalities for general measured length spaces. He received the

(c) visualization features

13th ACM Symposium on DocEng – Sept 10-13, Florence (Italy)









Thank for your attention !





References

-  Talel Abdessalem, M. Lamine Ba, and Pierre Senellart, *A probabilistic XML merging tool*, EDBT, 2011, Demonstration.
-  M. Lamine Ba, Talel Abdessalem, and Pierre Senellart, *Towards a version control model with uncertain data*, PIKM, 2011.
-  Evgeny Kharlamov, Werner Nutt, and Pierre Senellart, *Updating Probabilistic XML*, Updates in XML, 2010.
-  Benny Kimelfeld and Pierre Senellart, *Probabilistic XML: Models and complexity*, Advances in Probabilistic Databases for Uncertain Information Management (Zongmin Ma and Li Yan, eds.), Springer-Verlag, 2013.

