A Probabilistic XML Merging Tool

Talel Abdessalem Mouhamadou Lamine BA Pierre Senellart
Télécom ParisTech Université Cheikh Anta DIOP Télécom ParisTech
Paris, France Dakar, Senegal Paris, France

What this tool aim at...

- Representing the outcome of semi-structured documents integration as a probabilistic tree
- Evaluating the uncertainty (modeled as probability values) of the result of the merge
- Querying the probabilistic repository with a subset of the XPath query language

Application domain: Wikipedia revisions
The tool enables merging the revisions of a given Wikipedia page with:
→ an efficient evaluation of the uncertainty of the obtained result
→ an automatic management of conflicts.

Probabilistic Documents

Pr(e₁) = 0.7
Pr(e₂) = 0.6

P-Document

Corresponding possible documents and their probabilities

Merging of Wikipedia revisions

- A two-way tree merging technique for P-Documents
- Two steps: Matching of Revisions and Merging Matches

1. Matching of Revisions
Input: two revisions r_k−1 and r_k and their associated event formula.
Output:
- Deleted nodes: x: x ∈ r_k−1 and x has no match in r_k.
- Added nodes: x: x ∈ r_k and x has no match in r_k−1.
- Matched couples (x, y): x ∈ r_k−1 and y ∈ r_k match.

2. Merging Matches
- Deleted nodes:
  \[ f_{\text{new}}(x) = f_{\text{old}}(x) \land \neg f_k \]
- Matched couple:
  \[ f_{\text{new}}(x) = f_{\text{old}}(x) \]
- For added nodes:
  \[ f_{\text{new}}(x) = f_k \]
  or
  \[ f_{\text{new}}(x) = f_{\text{old}}(x) \lor f_k \]

The result of the merge process

Architecture of the system

System for managing Wikipedia documents.

Features

A keyword-based search engine for Wikipedia pages
Extracting the revisions of a given page
Selecting the list of revisions to merge
Building one’s own Wikipedia article
Displaying the result of the merge
Demonstrating a certain number of use cases
Using a subset of XPath query language