Understanding the Hidden Web

Pierre Senellart

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA FUTURS
UNIVERSITÉ PARIS-SUD 11

Max-Planck-Institut für Informatik, 22 August 2007

The Hidden Web

Definition (Hidden Web, Deep Web, Invisible Web)

The part of Web content not accessible from the hyperlinked structure of the World Wide Web. Typically: HTML forms, Web Services.

Size estimate (2001) : 500 times larger than the surface Web.

How to understand it and benefit from its content?

The Hidden Web

Definition (Hidden Web, Deep Web, Invisible Web)

The part of Web content not accessible from the hyperlinked structure of the World Wide Web. Typically: HTML forms, Web Services.

Size estimate (2001) : 500 times larger than the surface Web.

How to understand it and benefit from its content?

The Hidden Web

Definition (Hidden Web, Deep Web, Invisible Web)

The part of Web content not accessible from the hyperlinked structure of the World Wide Web. Typically: HTML forms, Web Services.

Size estimate (2001) : 500 times larger than the surface Web.

How to understand it and benefit from its content?

Understanding the Hidden Web

Purpose

- Intensional indexing of the Hidden Web.
- High-level queries.
- ⇒ a semantic search engine over the Hidden Web.

In a fully automatic, unsupervised, way!

- Difficult and broad problem.
- Use of domain knowledge (ontology, instances).
- Example of the database publication domain.

Understanding the Hidden Web

Purpose

- Intensional indexing of the Hidden Web.
- High-level queries.
- ⇒ a semantic search engine over the Hidden Web.

In a fully automatic, unsupervised, way!

- Difficult and broad problem.
- Use of domain knowledge (ontology, instances).
- Example of the database publication domain.

Understanding the Hidden Web

Purpose

- Intensional indexing of the Hidden Web.
- High-level queries.
- ⇒ a semantic search engine over the Hidden Web.

In a fully automatic, unsupervised, way!

- Difficult and broad problem.
- Use of domain knowledge (ontology, instances).
- Example of the database publication domain.

Understanding the Hidden Web

Purpose

- Intensional indexing of the Hidden Web.
- High-level queries.
- ⇒ a semantic search engine over the Hidden Web.

In a fully automatic, unsupervised, way!

- Difficult and broad problem.
- Use of domain knowledge (ontology, instances).
- Example of the database publication domain.

Understanding the Hidden Web

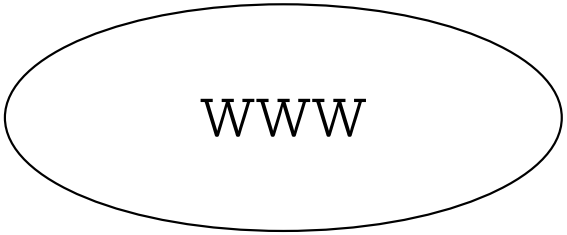
Purpose

- Intensional indexing of the Hidden Web.
- High-level queries.
- ⇒ a semantic search engine over the Hidden Web.

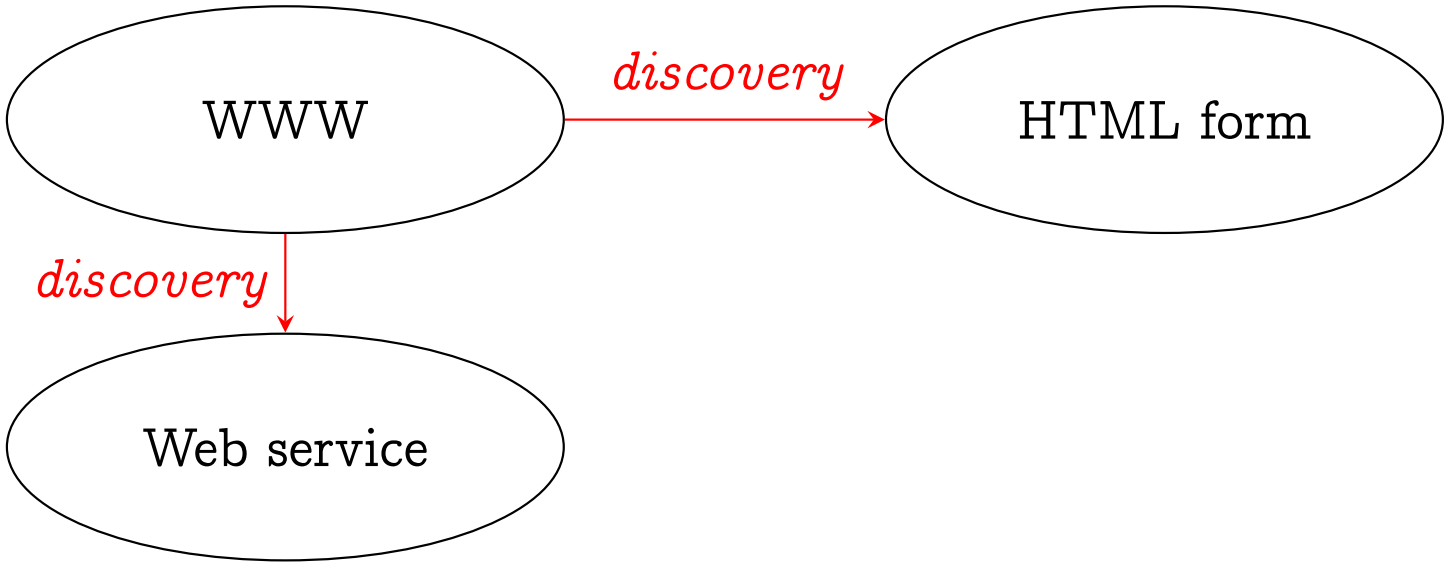
In a fully automatic, unsupervised, way!

- Difficult and broad problem.
- Use of domain knowledge (ontology, instances).
- Example of the database publication domain.

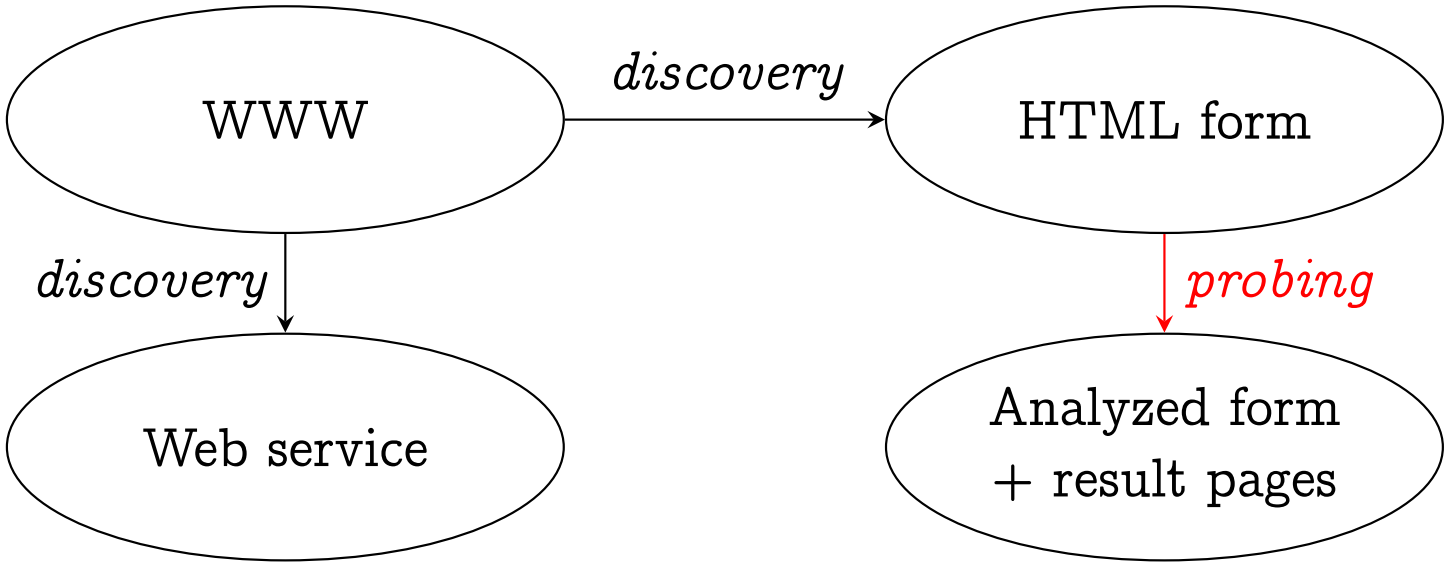
Web Service Semantic Interpretation Process



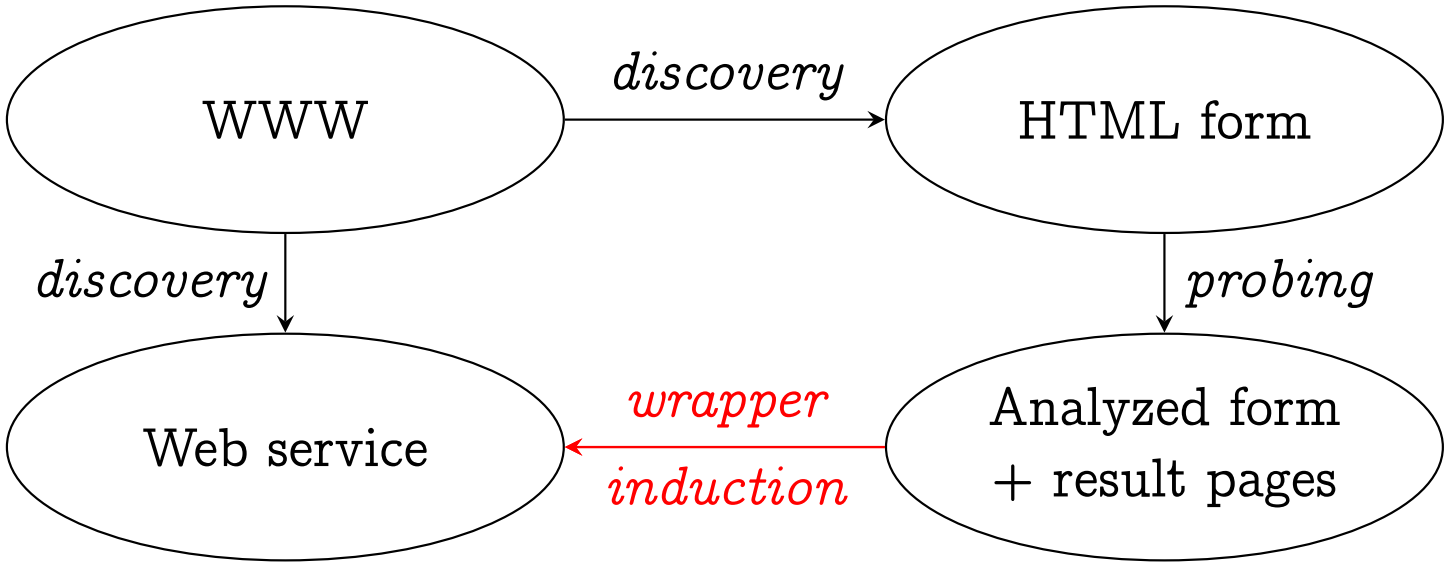
Web Service Semantic Interpretation Process



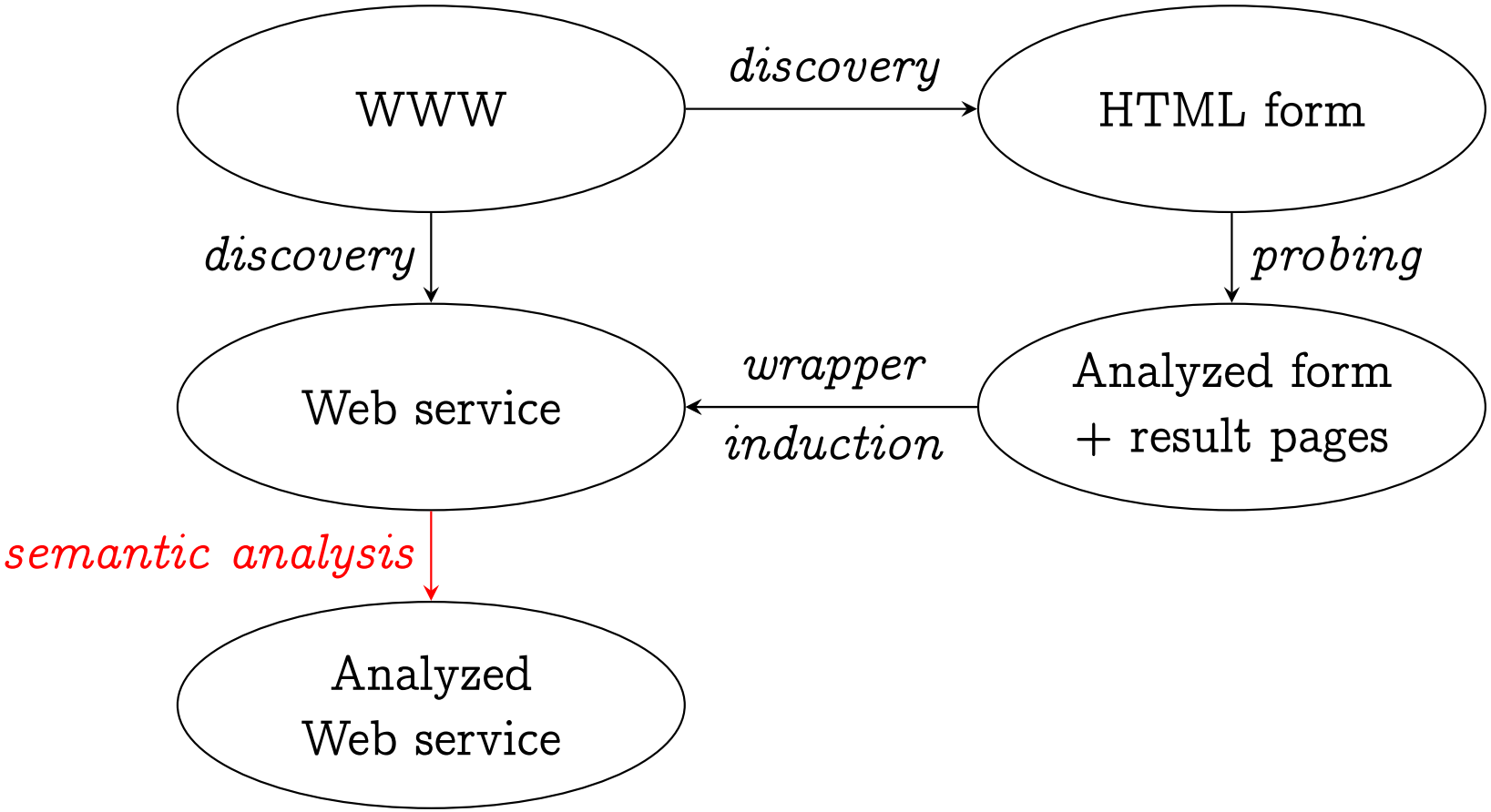
Web Service Semantic Interpretation Process



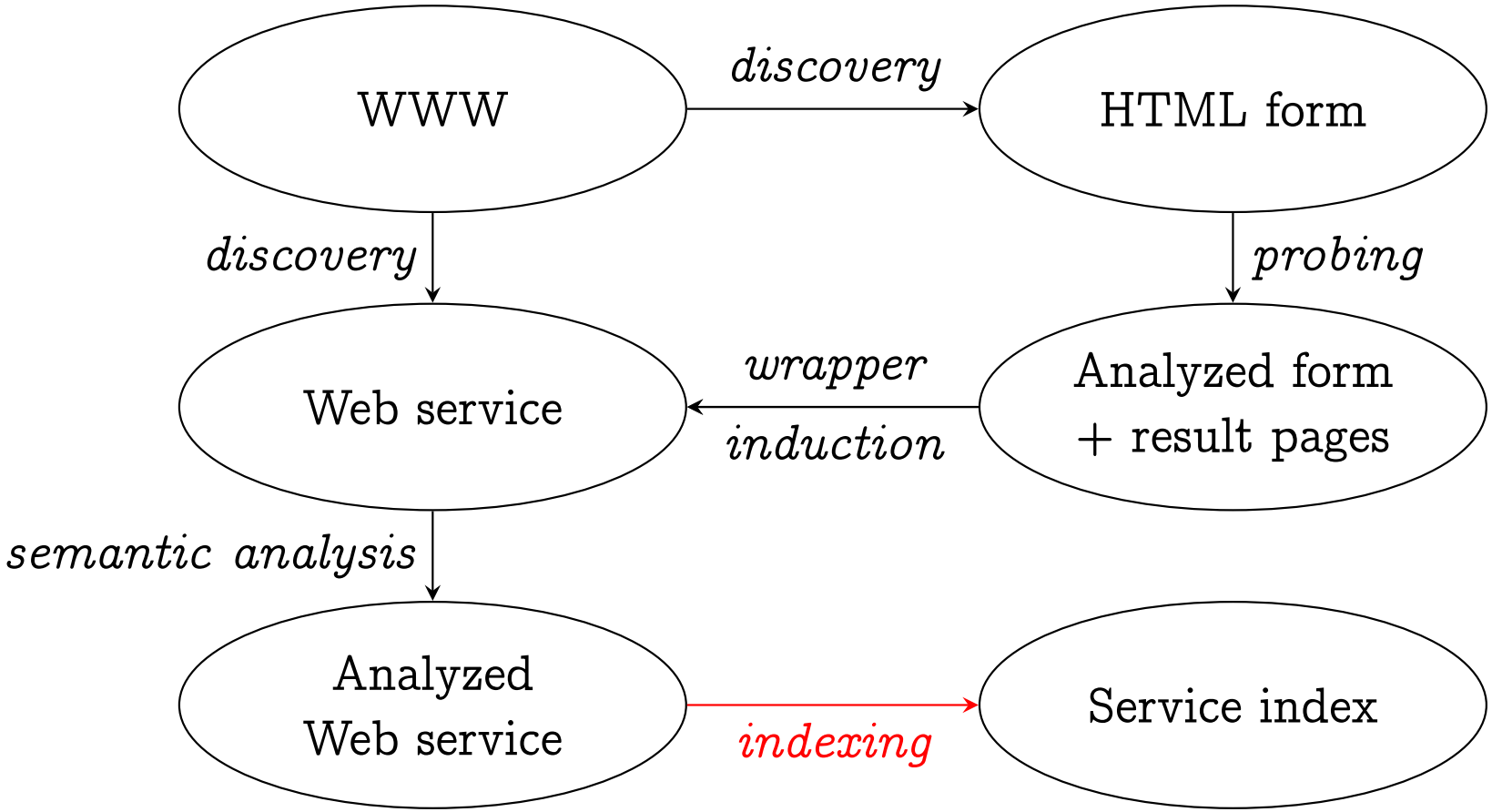
Web Service Semantic Interpretation Process



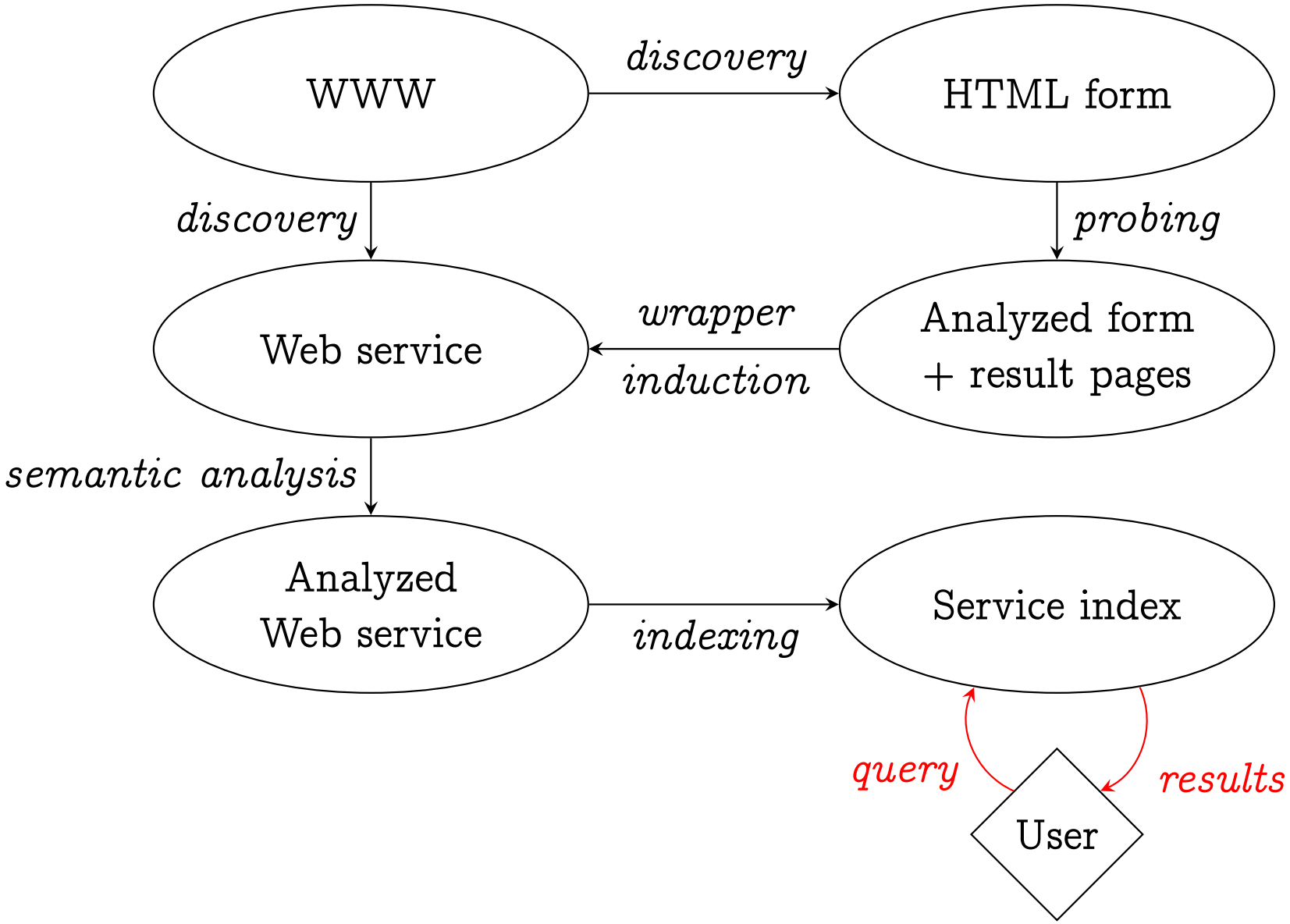
Web Service Semantic Interpretation Process



Web Service Semantic Interpretation Process



Web Service Semantic Interpretation Process



Imprecise Data and Imprecise Tasks

Observations

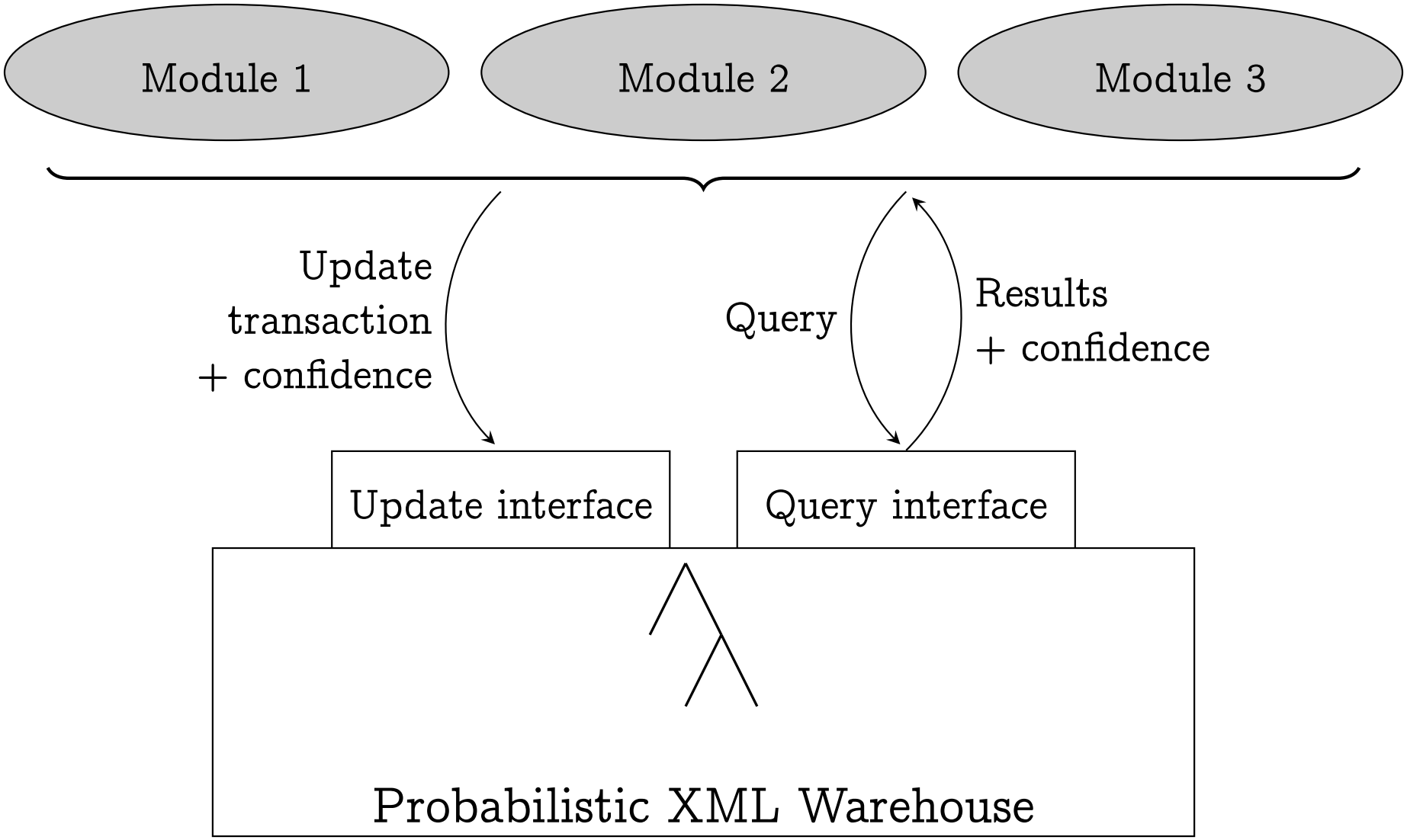
- Many needed tasks generate imprecise data, with some confidence value.
- Need for a way to manage this imprecision, to work with it throughout an entire complex process.

Imprecise Data and Imprecise Tasks

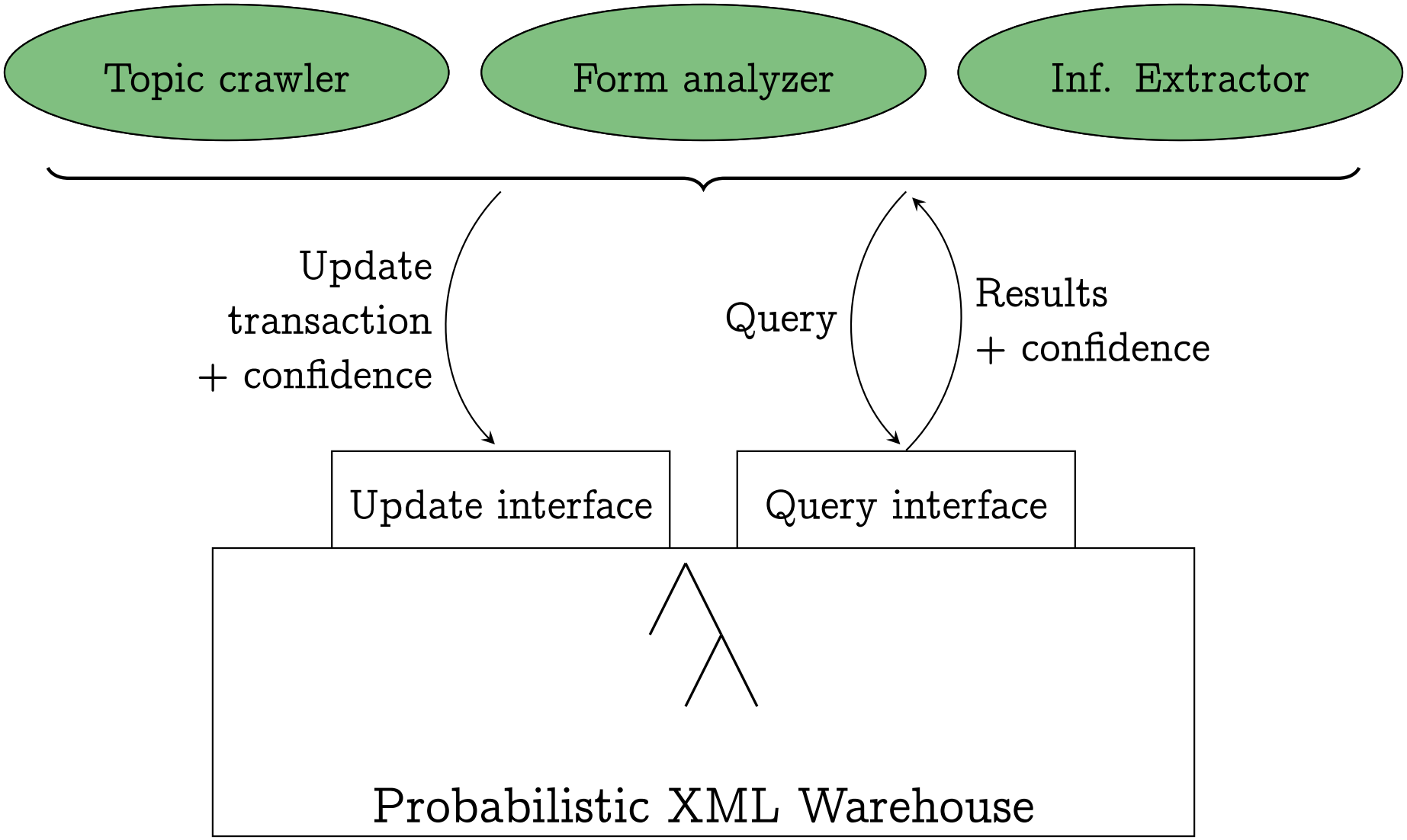
Observations

- Many needed tasks generate imprecise data, with some confidence value.
- Need for a way to manage this imprecision, to work with it throughout an entire complex process.

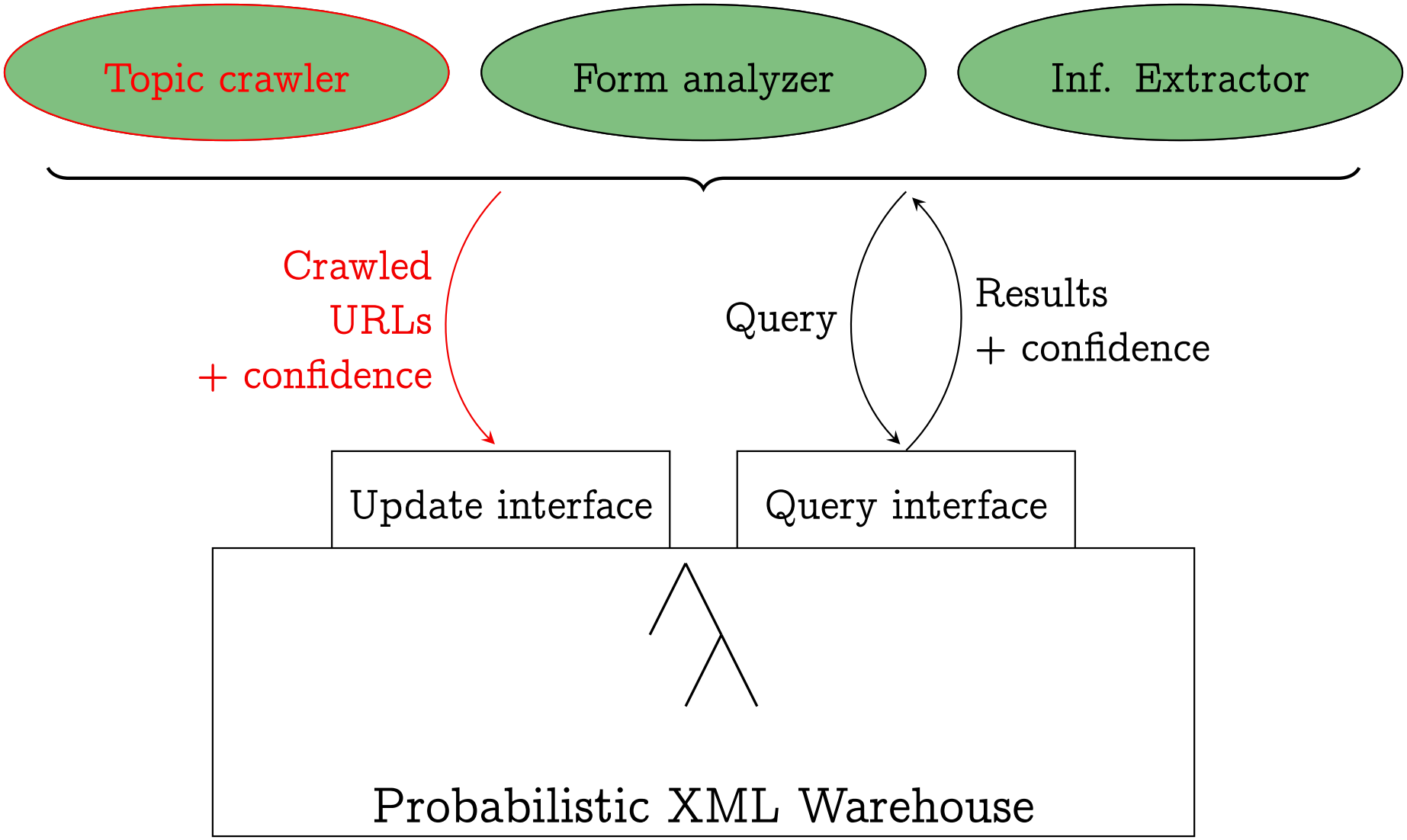
A Probabilistic XML Warehouse



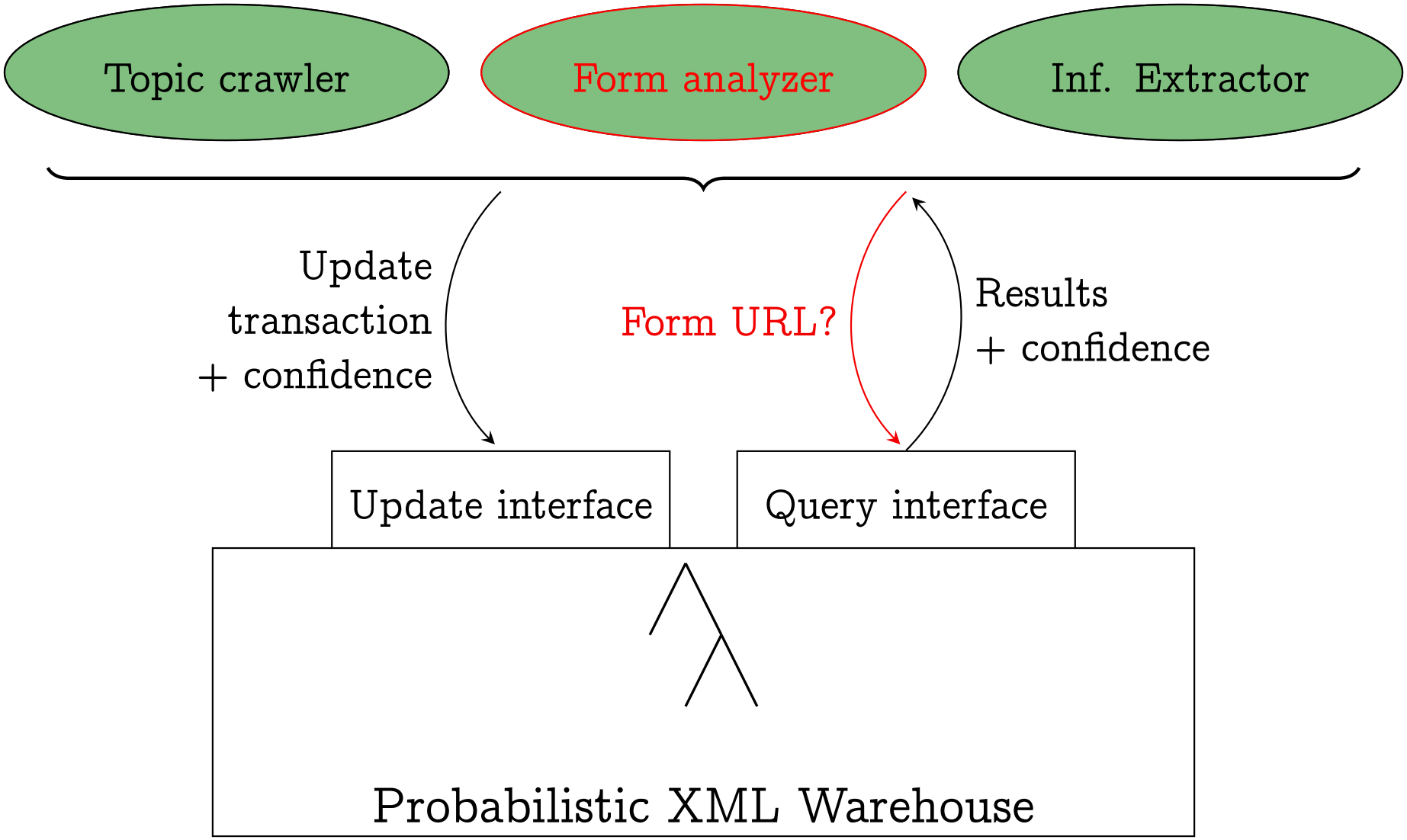
A Probabilistic XML Warehouse (Hidden Web)



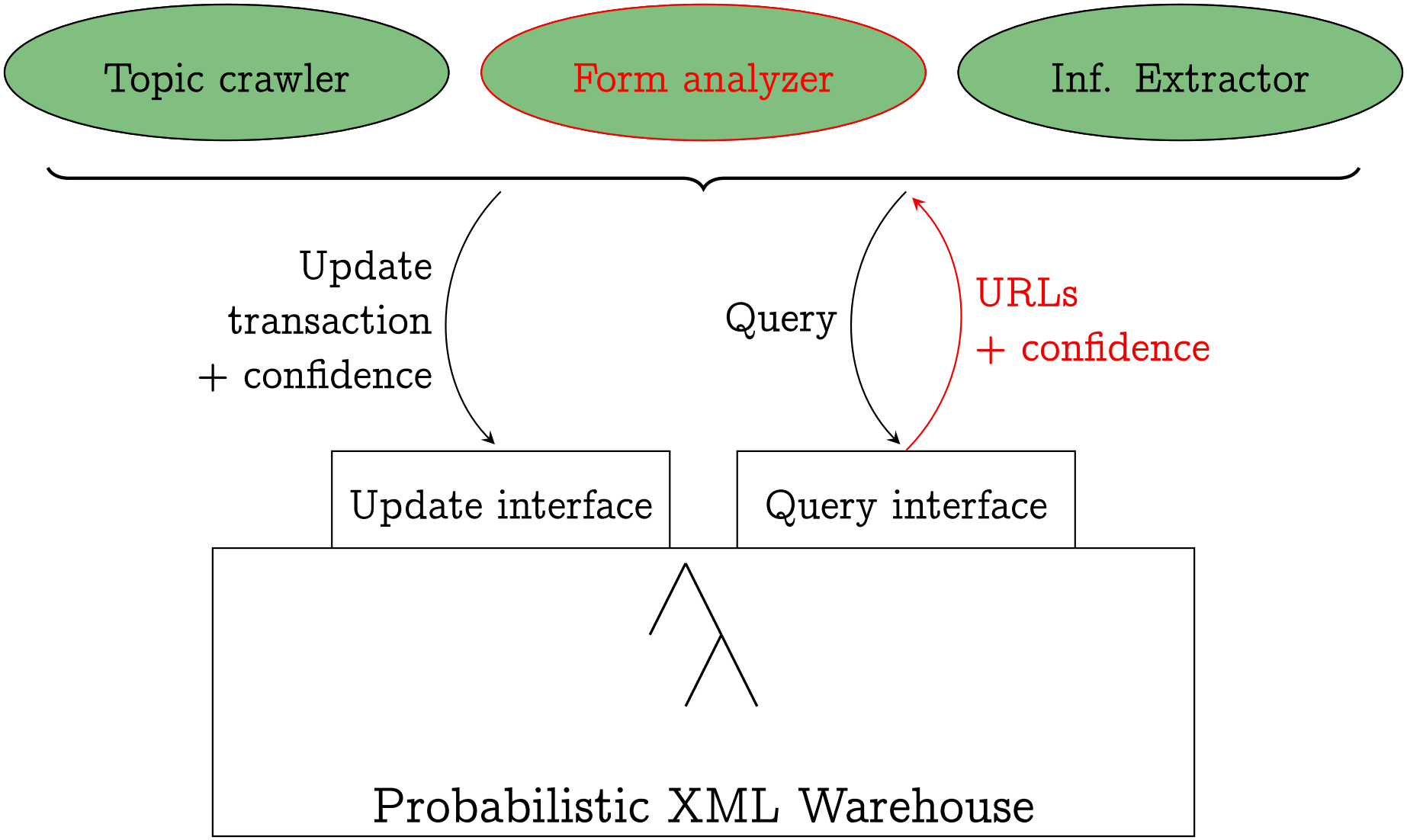
A Probabilistic XML Warehouse (Hidden Web)



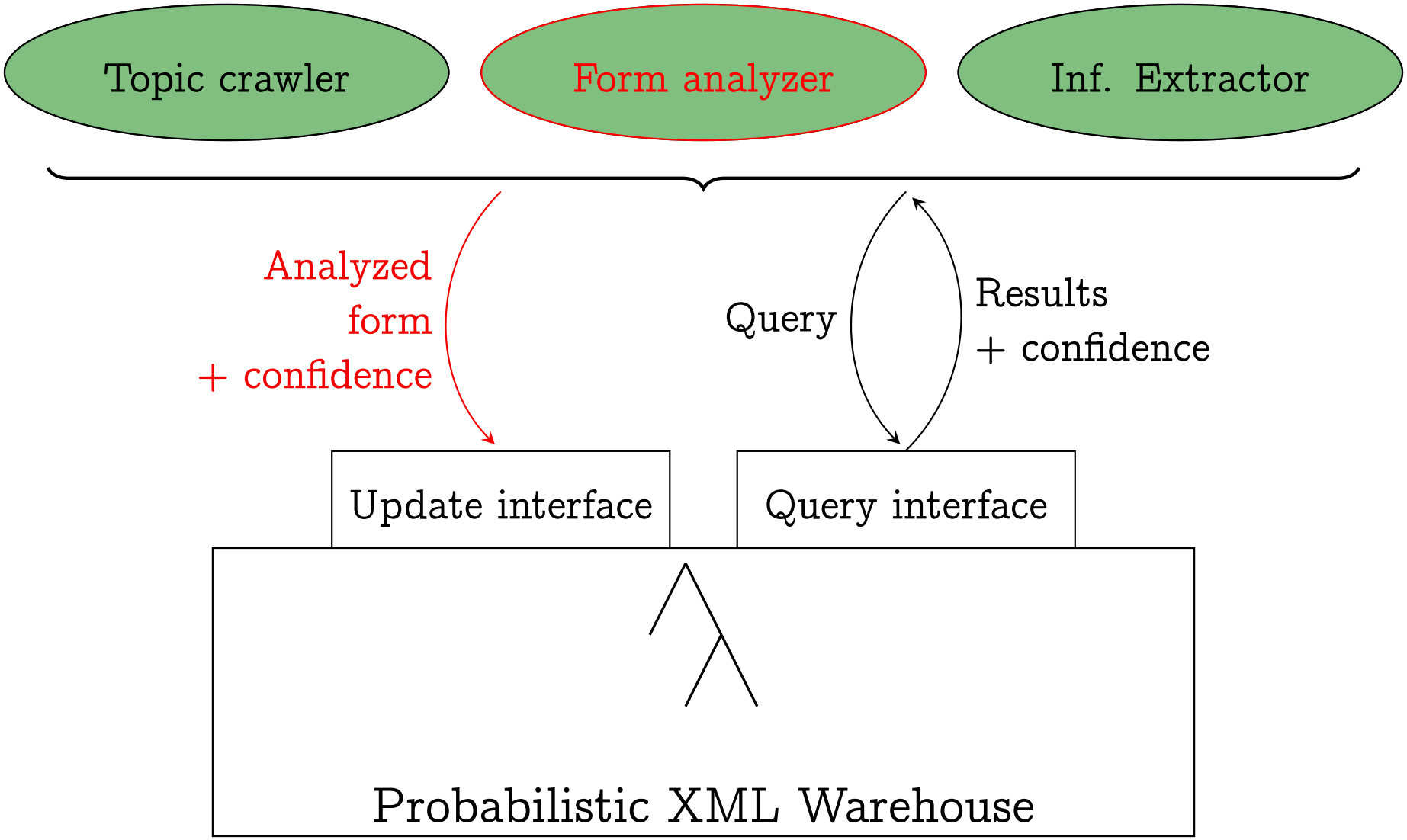
A Probabilistic XML Warehouse (Hidden Web)



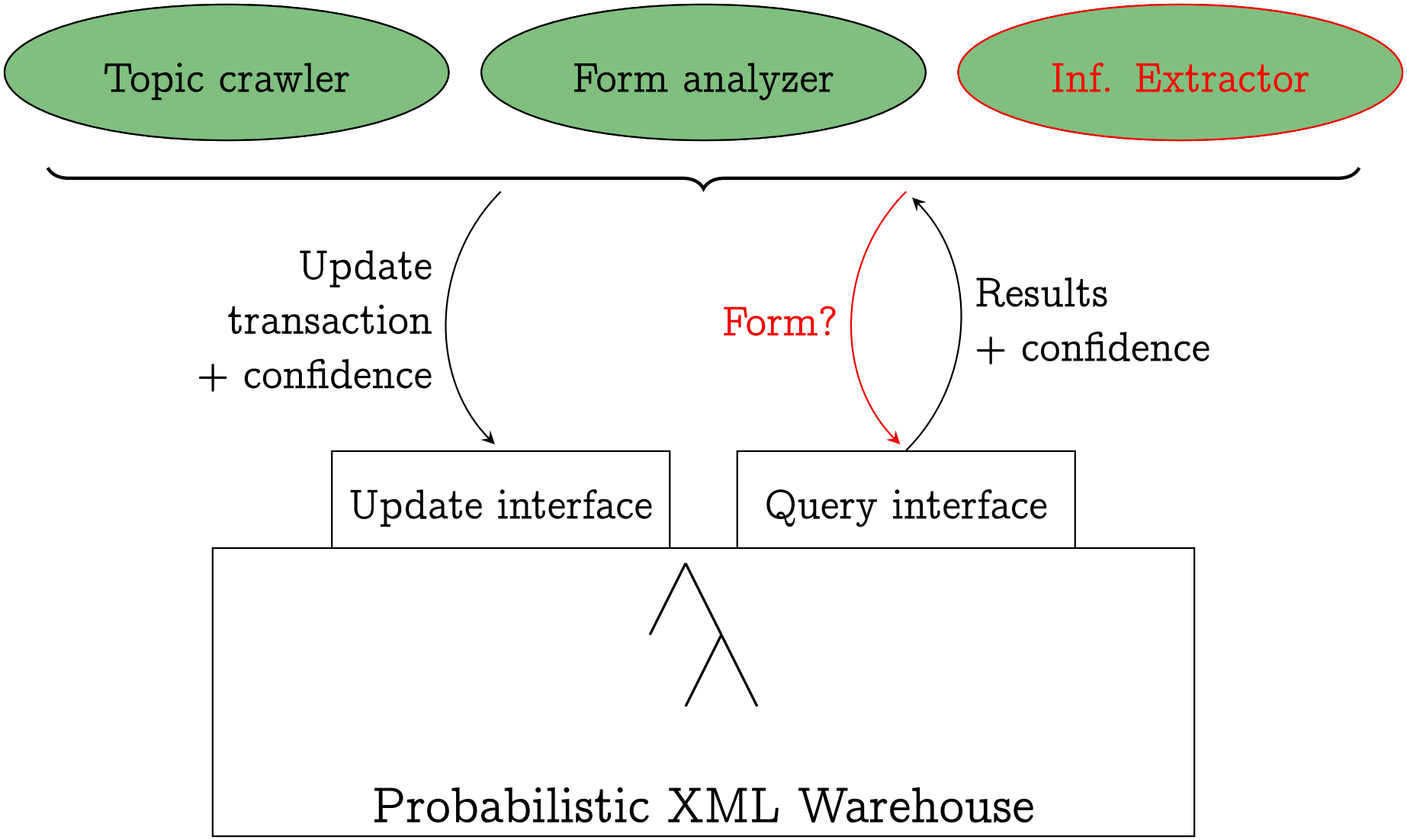
A Probabilistic XML Warehouse (Hidden Web)



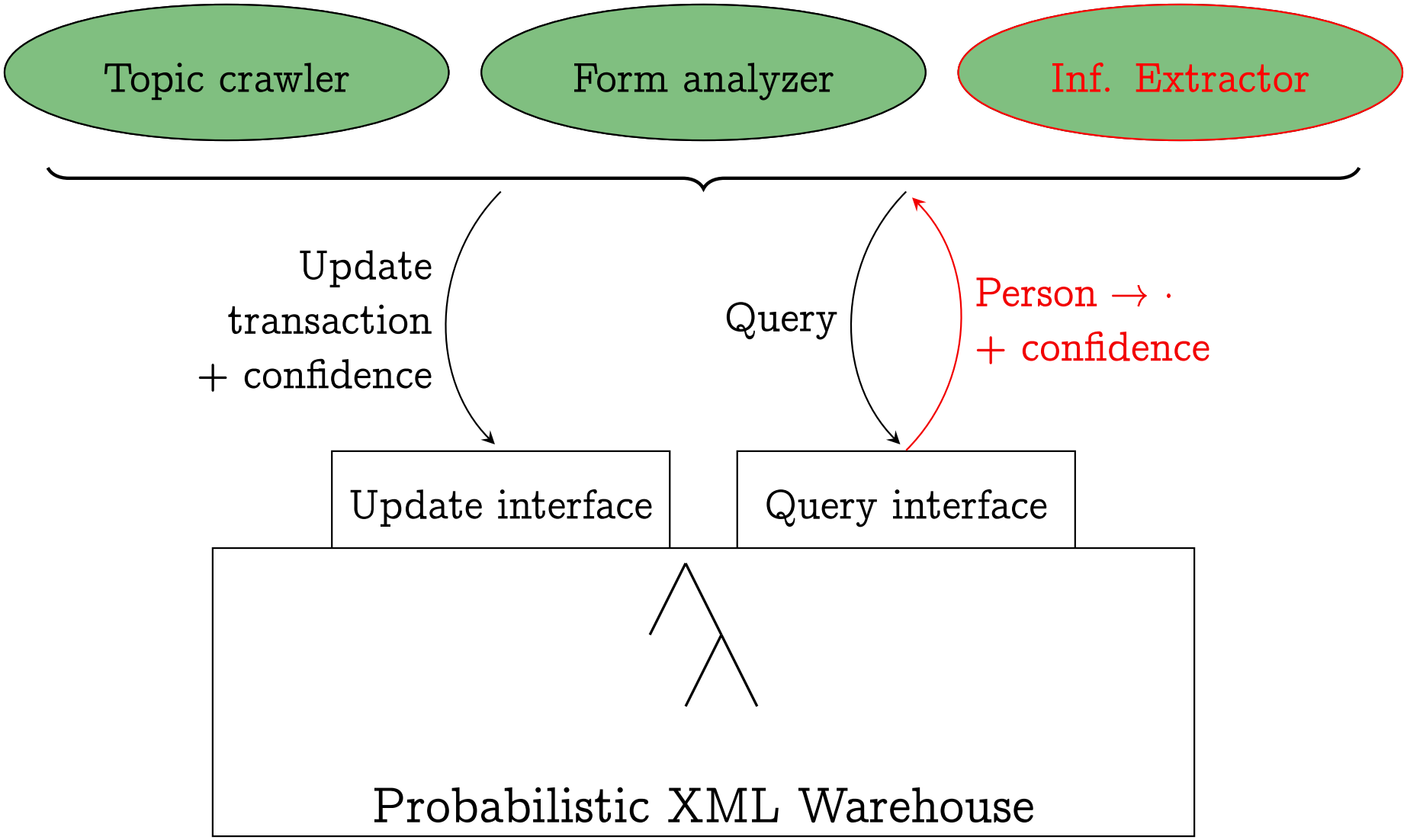
A Probabilistic XML Warehouse (Hidden Web)



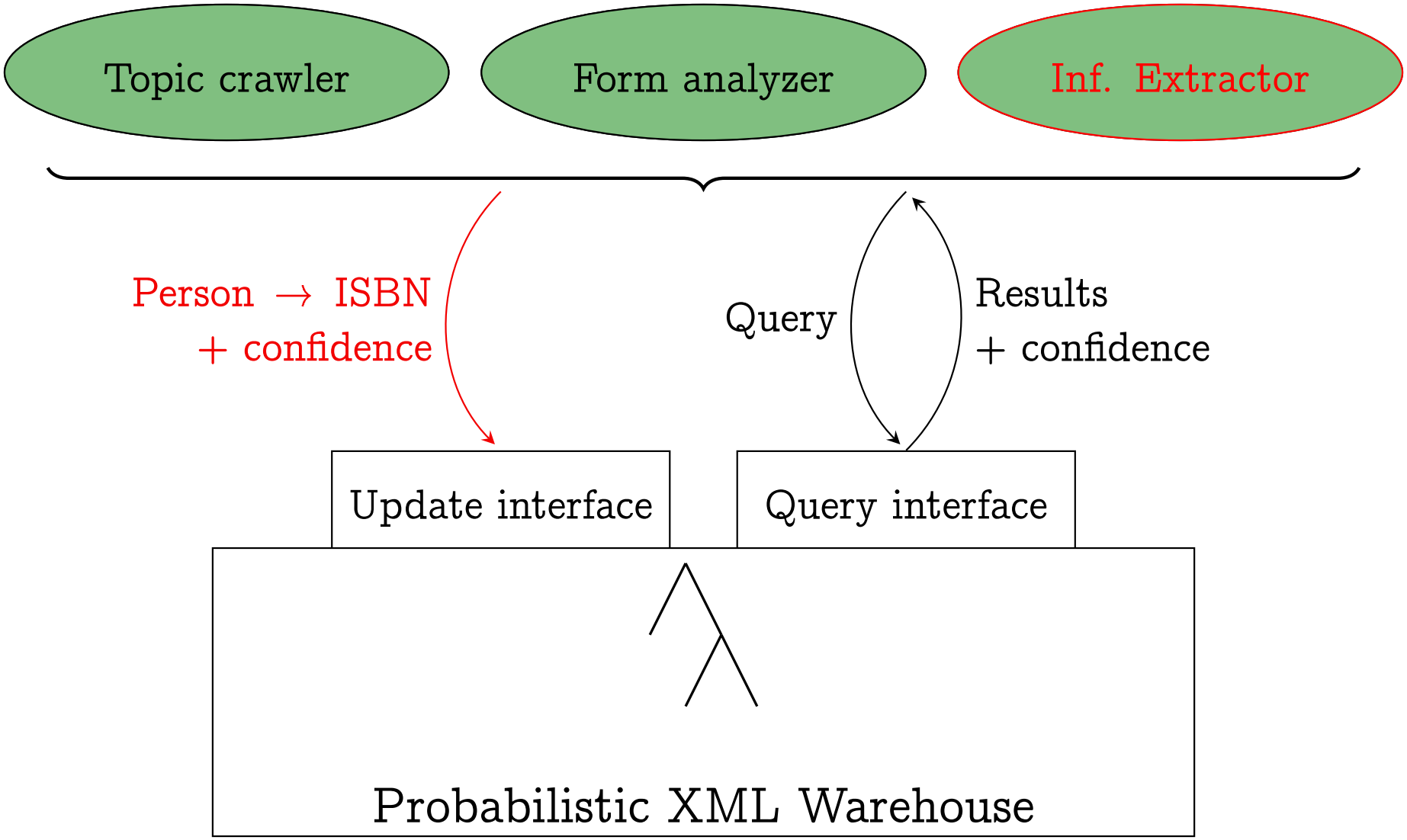
A Probabilistic XML Warehouse (Hidden Web)



A Probabilistic XML Warehouse (Hidden Web)



A Probabilistic XML Warehouse (Hidden Web)



Outline

Probabilistic Trees

Framework

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

(multiset semantics)

Sample space: Set of all such data trees.

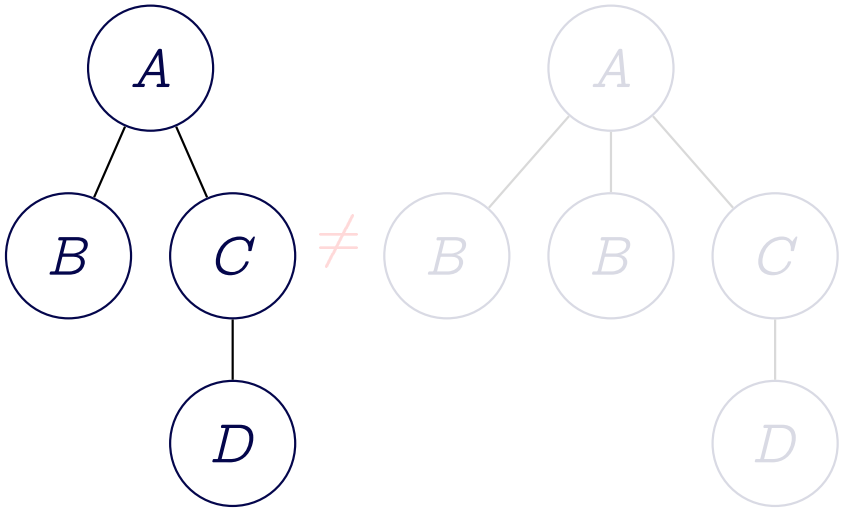
Probabilistic tree (prob-tree): Representation of a discrete probability distribution over this sample space.

Joint work with Serge Abiteboul.

Probabilistic Trees

Framework

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



(multiset semantics)

Sample space: Set of all such data trees.

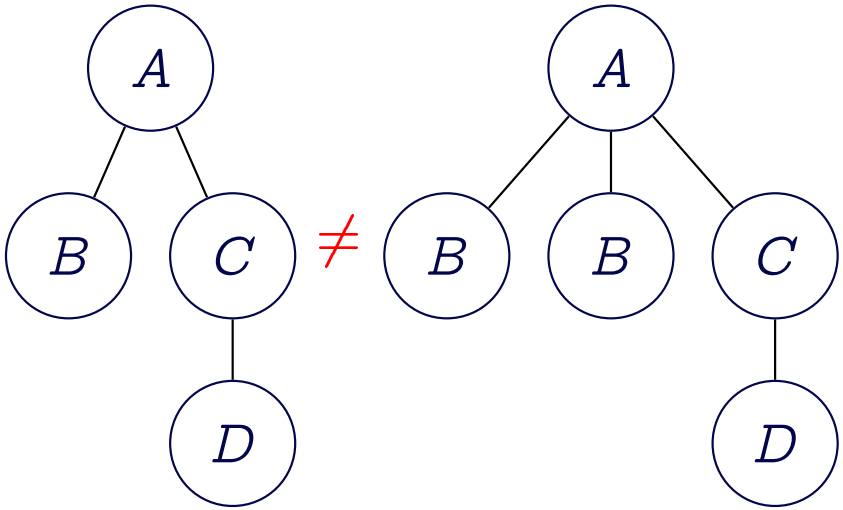
Probabilistic tree (prob-tree): Representation of a discrete probability distribution over this sample space.

Joint work with Serge Abiteboul.

Probabilistic Trees

Framework

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



(multiset semantics)

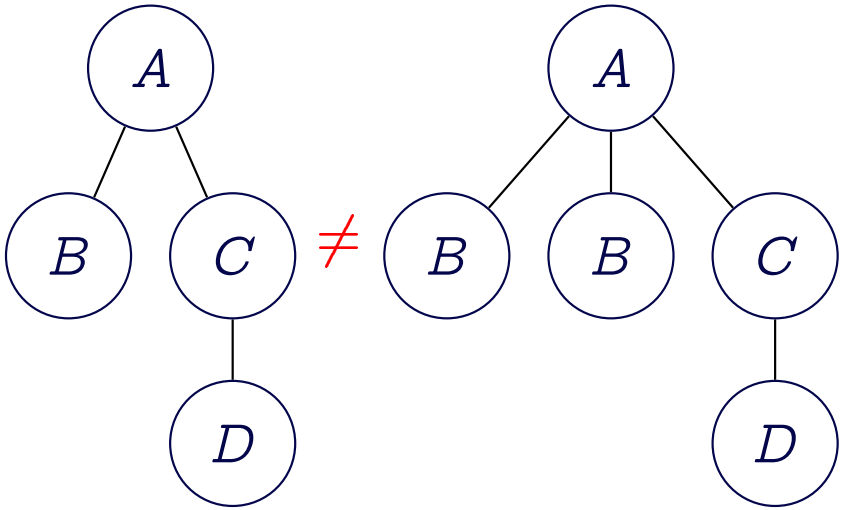
Sample space: Set of all such data trees.

Probabilistic tree (prob-tree): Representation of a discrete probability distribution over this sample space.

Joint work with Serge Abiteboul.

Probabilistic Trees

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



(multiset semantics)

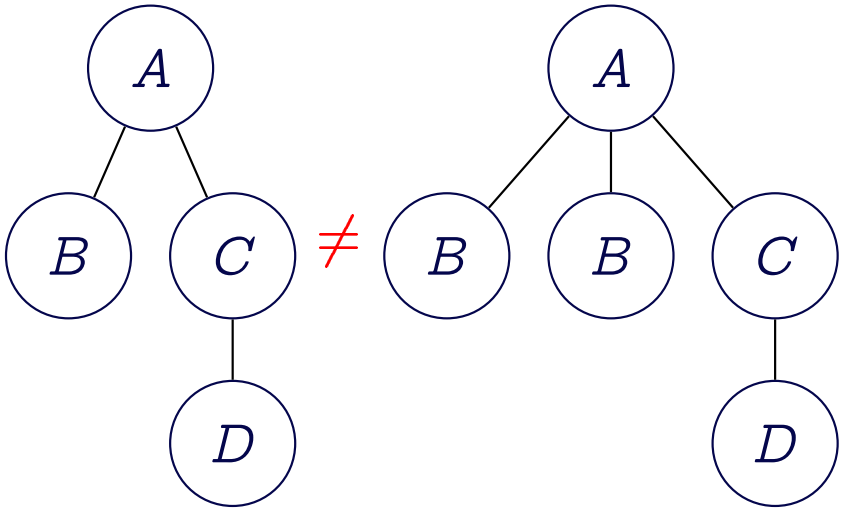
Sample space: Set of all such data trees.

Probabilistic tree (prob-tree): Representation of a discrete probability distribution over this sample space.

Joint work with Serge Abiteboul.

Probabilistic Trees

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



(multiset semantics)

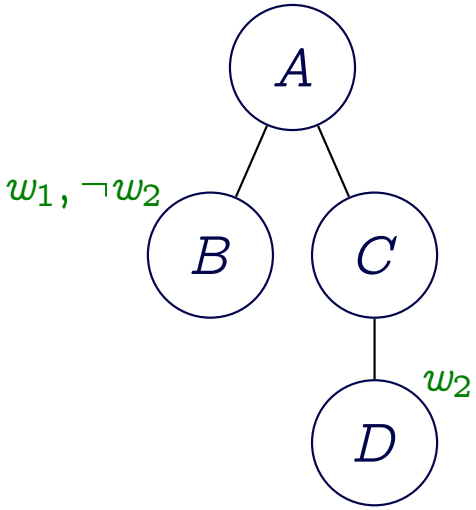
Sample space: Set of all such data trees.

Probabilistic tree (prob-tree): Representation of a discrete probability distribution over this sample space.

Joint work with Serge Abiteboul.

The Prob-Tree Model

- Data tree with event conditions (conjunction of probabilistic events or negations of probabilistic events) assigned to each node.
- Probabilistic events are boolean random variables, assumed to be independent, with their own probability distribution.



Event	Prob.
w_1	0.8
w_2	0.7

Joint work with Serge Abiteboul.

Features of the Prob-Tree Model

- Well-defined possible world semantics.
- Full expressive power, reasonable conciseness.
- Possible to apply query and updates directly on prob-trees, in an efficient way.
- Complexity study.
- Implementation available.

Joint work with Serge Abiteboul.

Features of the Prob-Tree Model

- Well-defined possible world semantics.
- Full expressive power, reasonable conciseness.
- Possible to apply query and updates directly on prob-trees, in an efficient way.
- Complexity study.
- Implementation available.

Joint work with Serge Abiteboul.

Features of the Prob-Tree Model

- Well-defined possible world semantics.
- Full expressive power, reasonable conciseness.
- Possible to apply query and updates directly on prob-trees, in an efficient way.
- Complexity study.
- Implementation available.

Joint work with Serge Abiteboul.

Features of the Prob-Tree Model

- Well-defined possible world semantics.
- Full expressive power, reasonable conciseness.
- Possible to apply query and updates directly on prob-trees, in an efficient way.
- Complexity study.
- Implementation available.

Joint work with Serge Abiteboul.

Features of the Prob-Tree Model

- Well-defined possible world semantics.
- Full expressive power, reasonable conciseness.
- Possible to apply query and updates directly on prob-trees, in an efficient way.
- Complexity study.
- Implementation available.

Joint work with Serge Abiteboul.

Outline

Analyzing HTML Forms

Analyzing the structure of HTML forms.

Authors				
Title			Year	Page
Conference			ID	
Journal			Volume	Number
Search	Reset	Maximum of 100 matches		

Problem

Associate to each relevant form field its corresponding domain concept.

Joint work with Avin Mittal (IIT Bombay).

First Step: Structural Analysis

1. Build a context for each field:
   - `label` tag;
   - `id` and `name` attributes;
   - text immediately before the field.
2. Remove stop words, stem.
3. Match this context with the concept names, extended with WordNet.
4. Obtain in this way candidate annotations.

Joint work with Avin Mittal (IIT Bombay).

First Step: Structural Analysis

1. Build a context for each field:
   - `label` tag;
   - `id` and `name` attributes;
   - text immediately before the field.
2. Remove stop words, stem.
3. Match this context with the concept names, extended with WordNet.
4. Obtain in this way candidate annotations.

Joint work with Avin Mittal (IIT Bombay).

First Step: Structural Analysis

1. Build a context for each field:
   - `label` tag;
   - `id` and `name` attributes;
   - text immediately before the field.
2. Remove stop words, stem.
3. Match this context with the concept names, extended with WordNet.
4. Obtain in this way candidate annotations.

Joint work with Avin Mittal (IIT Bombay).

First Step: Structural Analysis

1. Build a context for each field:
   - `label` tag;
   - `id` and `name` attributes;
   - text immediately before the field.
2. Remove stop words, stem.
3. Match this context with the concept names, extended with WordNet.
4. Obtain in this way candidate annotations.

Joint work with Avin Mittal (IIT Bombay).

Second Step: Confirm Annotations with Probing

For each field annotated with a concept c:

1. Probe the field with nonsense word to get an error page.
2. Probe the field with instances of c (chosen representatively of the frequency distribution of c).
3. Compare pages obtained by probing with the error page (by using clustering along the DOM tree structure of the pages), to distinguish error pages and result pages.
4. Confirm the annotation if enough result pages are obtained.

In practice, very good precision and good recall; but some limitations on the kind of forms that can be dealt with.

Joint work with Avin Mittal (IIT Bombay).

Second Step: Confirm Annotations with Probing

For each field annotated with a concept c:

1. Probe the field with nonsense word to get an error page.
2. Probe the field with instances of c (chosen representatively of the frequency distribution of c).
3. Compare pages obtained by probing with the error page (by using clustering along the DOM tree structure of the pages), to distinguish error pages and result pages.
4. Confirm the annotation if enough result pages are obtained.

In practice, very good precision and good recall; but some limitations on the kind of forms that can be dealt with.

Joint work with Avin Mittal (IIT Bombay).

Second Step: Confirm Annotations with Probing

For each field annotated with a concept c:

1. Probe the field with nonsense word to get an error page.
2. Probe the field with instances of c (chosen representatively of the frequency distribution of c).
3. Compare pages obtained by probing with the error page (by using clustering along the DOM tree structure of the pages), to distinguish error pages and result pages.
4. Confirm the annotation if enough result pages are obtained.

In practice, very good precision and good recall; but some limitations on the kind of forms that can be dealt with.

Joint work with Avin Mittal (IIT Bombay).

Second Step: Confirm Annotations with Probing

For each field annotated with a concept c:

1. Probe the field with nonsense word to get an error page.
2. Probe the field with instances of c (chosen representatively of the frequency distribution of c).
3. Compare pages obtained by probing with the error page (by using clustering along the DOM tree structure of the pages), to distinguish error pages and result pages.
4. Confirm the annotation if enough result pages are obtained.

In practice, very good precision and good recall; but some limitations on the kind of forms that can be dealt with.

Joint work with Avin Mittal (IIT Bombay).

Second Step: Confirm Annotations with Probing

For each field annotated with a concept c:

1. Probe the field with nonsense word to get an error page.
2. Probe the field with instances of c (chosen representatively of the frequency distribution of c).
3. Compare pages obtained by probing with the error page (by using clustering along the DOM tree structure of the pages), to distinguish error pages and result pages.
4. Confirm the annotation if enough result pages are obtained.

In practice, very good precision and good recall; but some limitations on the kind of forms that can be dealt with.

Joint work with Avin Mittal (IIT Bombay).

Outline

Query-answer Web Pages

Extract data from query-answer Web pages.

Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: **Riccardo Pucella, Vicky Weissman**
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: **Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)**
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: **Philip Bille, Inge Li Goertz**
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: **Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)**
Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: **Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)**
Subj-class: Cryptography and Security; Databases

Issues

- What part of the Web page contains the answer?
- How to extract structured content?

Joint work with researchers from MOSTRARE (INRIA Futurs).

Query-answer Web Pages

Extract data from query-answer Web pages.

Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: **Riccardo Pucella, Vicky Weissman**
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: **Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)**
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: **Philip Bille, Inge Li Goertz**
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: **Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)**
Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: **Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)**
Subj-class: Cryptography and Security; Databases

Issues

- What part of the Web page contains the answer?
- How to extract structured content?

Joint work with researchers from MOSTRARE (INRIA Futurs).

Automatic Wrapper Induction with Domain Knowledge

- Annotate pages with knowledge domain (finite automata techniques): Both imperfect and incomplete.

Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: Riccardo Pucella, Vicky Weissman
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: Philip Bille, Inge Li Goertz
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)
Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)
Subj-class: Cryptography and Security; Databases

- Use machine learning to generalize the result into a structural extraction wrapper (Conditional Random Fields).

Joint work with researchers from MOSTRARE (INRIA Futurs).

Automatic Wrapper Induction with Domain Knowledge

- Annotate pages with knowledge domain (finite automata techniques): Both imperfect and incomplete.

Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: Riccardo Pucella, Vicky Weissman
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: Philip Bille, Inge Li Goertz
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)
Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)
Subj-class: Cryptography and Security; Databases

- Use machine learning to generalize the result into a structural extraction wrapper (Conditional Random Fields).

Joint work with researchers from MOSTRARE (INRIA Futurs).

Outline

Motivation

Analyzing the relations between different sources, or between a source and the domain knowledge.

Problem

Given two database instances I and J with different schemata, what is the optimal description Σ of J with respect to I (with Σ a finite set of formulæ in some logical language)?

What does optimal implies:

- Conciseness of description.
- Validity of facts predicted by I and Σ.
- Facts of J explained by I and Σ.

(Note the asymmetry between I and J; context of data exchange where J is computed from I and Σ).

Joint work with Georg Gottlob (Oxford University).

Motivation

Analyzing the relations between different sources, or between a source and the domain knowledge.

Problem

Given two database instances I and J with different schemata, what is the optimal description Σ of J with respect to I (with Σ a finite set of formulæ in some logical language)?

What does optimal implies:

- Conciseness of description.
- Validity of facts predicted by I and Σ.
- Facts of J explained by I and Σ.

(Note the asymmetry between I and J; context of data exchange where J is computed from I and Σ).

Joint work with Georg Gottlob (Oxford University).

Motivation

Analyzing the relations between different sources, or between a source and the domain knowledge.

Problem

Given two database instances I and J with different schemata, what is the optimal description Σ of J with respect to I (with Σ a finite set of formulæ in some logical language)?

What does optimal implies:

- Conciseness of description.
- Validity of facts predicted by I and Σ.
- Facts of J explained by I and Σ.

(Note the asymmetry between I and J; context of data exchange where J is computed from I and Σ).

Joint work with Georg Gottlob (Oxford University).

Example (Tuple-Generating Dependencies)

R
a
b
c
d

R'	
a	a
b	b
c	a
d	d
g	h

$$\Sigma_0 = \varnothing$$

$$\Sigma_1 = \{\forall x\ R(x) \rightarrow R'(x,x)\}$$

$$\Sigma_2 = \{\forall x\ R(x) \rightarrow \exists y\ R'(x,y)\}$$

$$\Sigma_3 = \{\forall x \forall y\ R(x) \wedge R(y) \rightarrow R'(x,y)\}$$

$$\Sigma_4 = \{\exists x \exists y\ R'(x,y)\}$$

Joint work with Georg Gottlob (Oxford University).

Results

- Description based on the minimum length of a repair of a formula that is valid and explains all facts of J.
- This optimality notion gives "intuitive" results for instances derived from each other with simple operations.
- Detailed complexity analysis for various languages and decision problems. Quite high in the polynomial hierarchy (up to Π_4^P for general tgds!).
- Even for $\forall x_1 \forall x_2 \forall x_3\ R(x_1, x_2, x_3) \rightarrow R'(x_1)$, computing the size of the minimal perfect repair is already NP-complete.

Joint work with Georg Gottlob (Oxford University).

Results

- Description based on the minimum length of a repair of a formula that is valid and explains all facts of J.
- This optimality notion gives "intuitive" results for instances derived from each other with simple operations.
- Detailed complexity analysis for various languages and decision problems. Quite high in the polynomial hierarchy (up to Π_4^P for general tgds!).
- Even for $\forall x_1 \forall x_2 \forall x_3\ R(x_1, x_2, x_3) \rightarrow R'(x_1)$, computing the size of the minimal perfect repair is already NP-complete.

Joint work with Georg Gottlob (Oxford University).

Results

- Description based on the minimum length of a repair of a formula that is valid and explains all facts of J.
- This optimality notion gives "intuitive" results for instances derived from each other with simple operations.
- Detailed complexity analysis for various languages and decision problems. Quite high in the polynomial hierarchy (up to Π_4^P for general tgds!).
- Even for $\forall x_1 \forall x_2 \forall x_3\ R(x_1, x_2, x_3) \rightarrow R'(x_1)$, computing the size of the minimal perfect repair is already NP-complete.

Joint work with Georg Gottlob (Oxford University).

Results

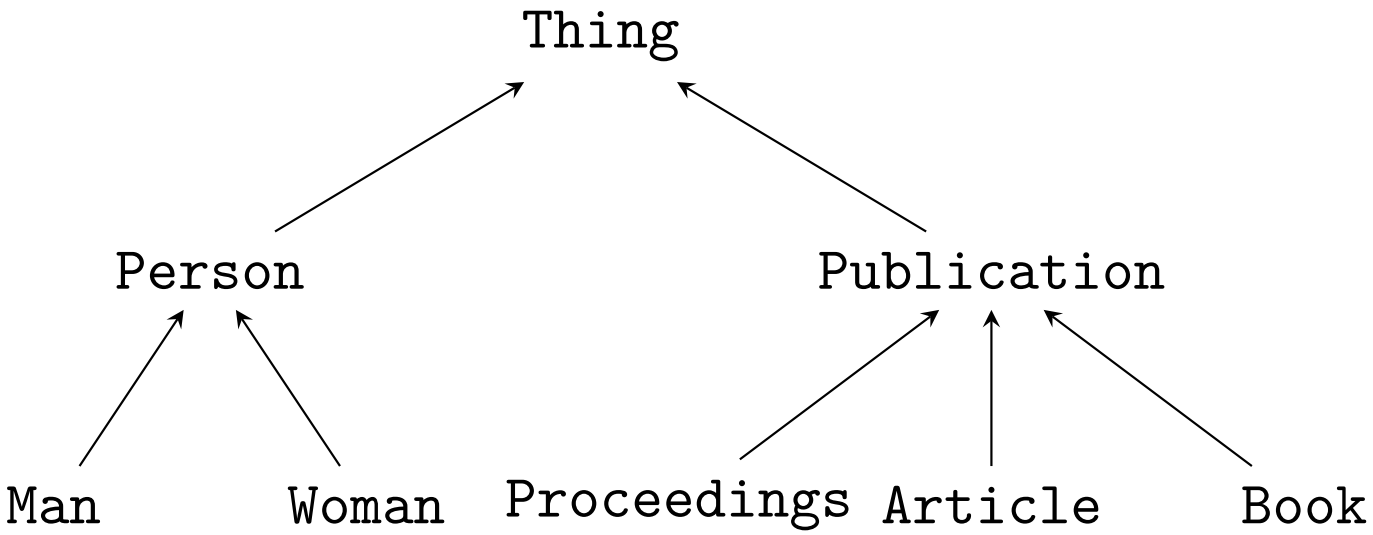
- Description based on the minimum length of a repair of a formula that is valid and explains all facts of J.
- This optimality notion gives "intuitive" results for instances derived from each other with simple operations.
- Detailed complexity analysis for various languages and decision problems. Quite high in the polynomial hierarchy (up to Π_4^P for general tgds!).
- Even for $\forall x_1 \forall x_2 \forall x_3\ R(x_1, x_2, x_3) \rightarrow R'(x_1)$, computing the size of the minimal perfect repair is already **NP-complete**.

Joint work with Georg Gottlob (Oxford University).

Outline

Conceptual Model

- IsA ontology of concepts (simple DAG)

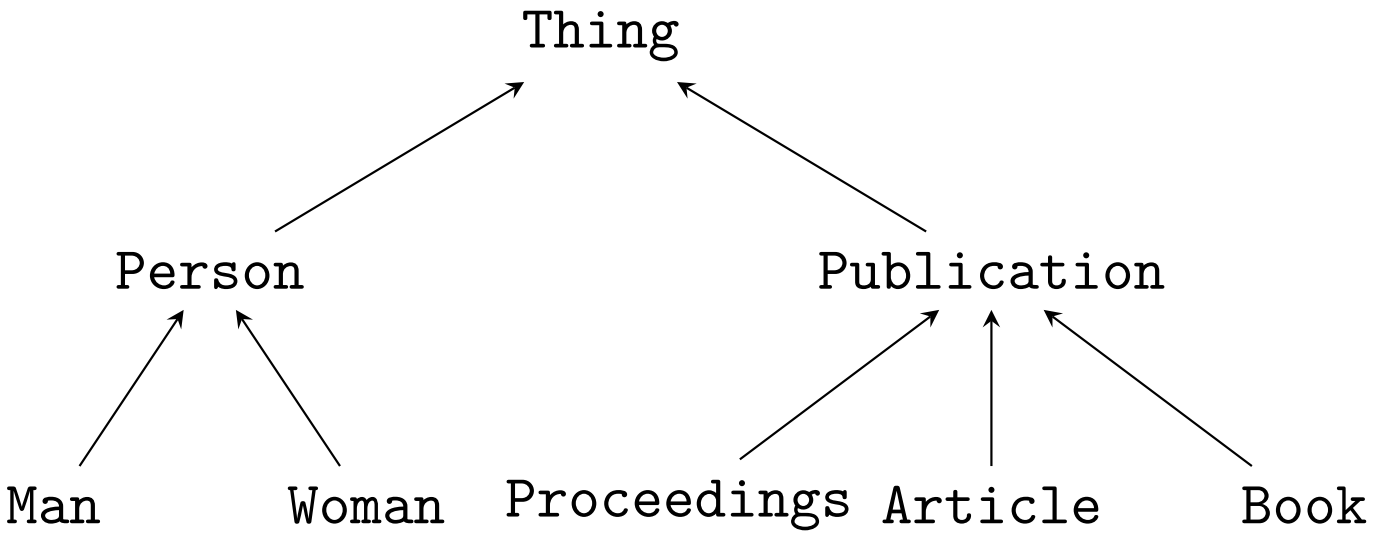


- n-ary typed roles
  - AuthorOf(Publication,Person)
  - HasName(Person,Name)

Joint work with Serge Abiteboul.

Conceptual Model

- IsA ontology of concepts (simple DAG)



- n-ary typed roles
  - `AuthorOf(Publication,Person)`
  - `HasName(Person,Name)`

Joint work with Serge Abiteboul.

Semantic Representation of a Service

What is a service described by?

- A n-uple of typed input parameters.
- A complex (= nested) type of its output.
- Semantic relations between inputs and outputs (Datalog-like description).

Definition (Complex types)

S: set of concepts

$$T \longleftarrow S | < T, \ldots, T > | T*$$

Joint work with Serge Abiteboul.

Semantic Representation of a Service

What is a service described by?

- A n-uple of typed input parameters.
- A complex (= nested) type of its output.
- Semantic relations between inputs and outputs (Datalog-like description).

Definition (Complex types)

S: set of concepts

$$T \longleftarrow S | < T, \ldots, T > | T*$$

Joint work with Serge Abiteboul.

Semantic Representation of a Service

What is a service described by?

- A n-uple of typed input parameters.
- A complex (= nested) type of its output.
- Semantic relations between inputs and outputs (Datalog-like description).

Definition (Complex types)

S: set of concepts

$$T \longleftarrow S | < T, \ldots, T > | T*$$

Joint work with Serge Abiteboul.

Semantic Representation of a Service

What is a service described by?

- A n-uple of typed input parameters.
- A complex (= nested) type of its output.
- Semantic relations between inputs and outputs (Datalog-like description).

Definition (Complex types)

S: set of concepts

$$T \longleftarrow S | < T, \ldots, T > | T*$$

Joint work with Serge Abiteboul.

Services and Queries

Example

Service giving authors from publication titles

$$A^* \leftarrow \texttt{AuthorOf}(A,P),\texttt{HasTitle}(P,T),\texttt{Input}(T)$$

Example

Query:

$$<A,T^*>^* \leftarrow \texttt{AuthorOf}(A,P),\ \texttt{Article}(P),\ \texttt{HasTitle}(P,T),\ \texttt{KeywordOf}(\text{“xml”},P)$$

Joint work with Serge Abiteboul.

Services and Queries

Example

Service giving authors from publication titles

A*← AuthorOf(A,P),HasTitle(P,T),Input(T)

Example

Query:

<A,T*>* ← AuthorOf(A,P), Article(P),
HasTitle(P,T), KeywordOf("xml",P)

Joint work with Serge Abiteboul.

Managing Extensional Information

How to represent extensional information (i.e. documents) in this formalism?

Definition

A document is a service with no input.

Complex types: natural representation of a DTD.

(Disjunctions `a|b` simulated by `(a?,b?)`).

Joint work with Serge Abiteboul.

Web Service Indexing and Querying

Given a query, represented as an analyzed Web service, how to know which known Web services to query?

Issues

- Subsumption of input/output parameters.
- Missing input parameters.
- Composition of Web Services.

Joint work with Serge Abiteboul.

Web Service Indexing and Querying

Given a query, represented as an analyzed Web service, how to know which known Web services to query?

Issues

- Subsumption of input/output parameters.
- Missing input parameters.
- Composition of Web Services.

Joint work with Serge Abiteboul.

Web Service Indexing and Querying

Given a query, represented as an analyzed Web service, how to know which known Web services to query?

Issues

- Subsumption of input/output parameters.
- Missing input parameters.
- Composition of Web Services.

Joint work with Serge Abiteboul.

Web Service Indexing and Querying

Given a query, represented as an analyzed Web service, how to know which known Web services to query?

Issues

- Subsumption of input/output parameters.
- Missing input parameters.
- Composition of Web Services.

Joint work with Serge Abiteboul.

Differences with Classical Database Querying

Three main differences:

- Information can be queried only through *views* (Local As View).
- Nested types.
- Incomplete information.

Three sources of complexity!

Current direction of work: Using Magic sets techniques (for evaluation of Datalog programs) restricted to appropriate binding patterns.

Joint work with Serge Abiteboul.

Differences with Classical Database Querying

Three main differences:

- Information can be queried only through *views* (Local As View).
- Nested types.
- Incomplete information.

Three sources of complexity!

Current direction of work: Using Magic sets techniques (for evaluation of Datalog programs) restricted to appropriate binding patterns.

Joint work with Serge Abiteboul.

Differences with Classical Database Querying

Three main differences:

- Information can be queried only through *views* (Local As View).
- Nested types.
- Incomplete information.

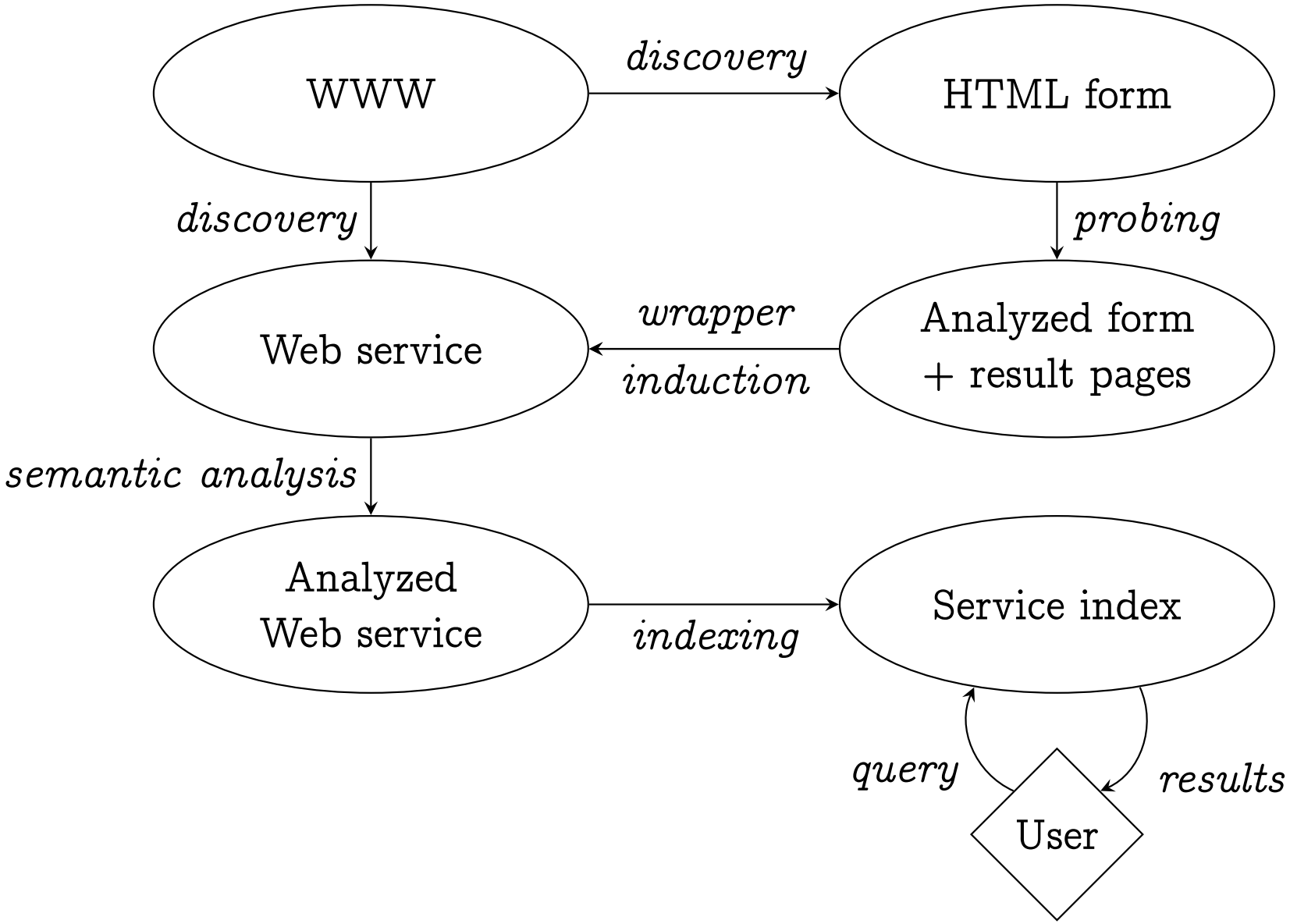
Three sources of complexity!

Current direction of work: Using Magic sets techniques (for evaluation of Datalog programs) restricted to appropriate binding patterns.

Joint work with Serge Abiteboul.

Outline

Web Service Semantic Interpretation Process



Perspectives

Still a lot to do... In particular:

- Answering queries using views on the semantic model.
- Continue work on automatic wrapper induction, to get a form fully wrapped as a Web service.
- Relation between schema mapping induction and inductive logic programming.

Other Works

Data Warehousing Extraction of information from the Web, mailing lists... to build a warehouse of sociological data (with various people).

Graph, Text and Web Mining

- Similarity between nodes in graphs; application to synonym extraction (with Vincent Blondel, from UCL).
- Related nodes in a graph; application to Wikipedia (with Yann Ollivier, from ÉNS Lyon).
- PageRank prediction (with Michalis Vazirgiannis).

Machine Translation Close relations with SYSTRAN; XML document processing, statistical and rule-based machine translation, multilingual authoring...

Other Works

Data Warehousing Extraction of information from the Web, mailing lists... to build a warehouse of sociological data (with various people).

Graph, Text and Web Mining

- Similarity between nodes in graphs; application to synonym extraction (with Vincent Blondel, from UCL).
- Related nodes in a graph; application to Wikipedia (with Yann Ollivier, from ÉNS Lyon).
- PageRank prediction (with Michalis Vazirgiannis).

Machine Translation Close relations with SYSTRAN; XML document processing, statistical and rule-based machine translation, multilingual authoring...

Other Works

Data Warehousing Extraction of information from the Web, mailing lists... to build a warehouse of sociological data (with various people).

Graph, Text and Web Mining

- Similarity between nodes in graphs; application to synonym extraction (with Vincent Blondel, from UCL).
- Related nodes in a graph; application to Wikipedia (with Yann Ollivier, from ÉNS Lyon).
- PageRank prediction (with Michalis Vazirgiannis).

Machine Translation Close relations with SYSTRAN; XML document processing, statistical and rule-based machine translation, multilingual authoring...

Other Works

Data Warehousing Extraction of information from the Web, mailing lists... to build a warehouse of sociological data (with various people).

Graph, Text and Web Mining

- Similarity between nodes in graphs; application to synonym extraction (with Vincent Blondel, from UCL).
- Related nodes in a graph; application to Wikipedia (with Yann Ollivier, from ÉNS Lyon).
- PageRank prediction (with Michalis Vazirgiannis).

Machine Translation Close relations with SYSTRAN; XML document processing, statistical and rule-based machine translation, multilingual authoring...

Other Works

Data Warehousing Extraction of information from the Web, mailing lists... to build a warehouse of sociological data (with various people).

Graph, Text and Web Mining

- Similarity between nodes in graphs; application to synonym extraction (with Vincent Blondel, from UCL).
- Related nodes in a graph; application to Wikipedia (with Yann Ollivier, from ÉNS Lyon).
- PageRank prediction (with Michalis Vazirgiannis).

Machine Translation Close relations with SYSTRAN; XML document processing, statistical and rule-based machine translation, multilingual authoring...