Understanding the Hidden Web

Pierre Senellart

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

TELECOM ParisTech

Athens University of Economics & Business, 28 July 2008

Simple problem

Contact all co-authors of Serge Abiteboul.

It's easy! You just need to:

- Find all co-authors.
- For each of them, find their current email address.

Google Scholar BETA

Advanced Scholar Search

Advanced Search Tips | About Google Scholar

Find articles with **all** of the words

with the **exact phrase**

with **at least one** of the words

without the words

where my words occur: anywhere in the article

10 results | Search Scholar

Author Return articles written by

e.g., *"PJ Hayes"* or *McCarthy*

Publication Return articles published in

e.g., *J Biol Chem* or *Nature*

Date Return articles published between —

e.g., *1996*

Subject Areas

- (●) Return articles in all subject areas.
- ( ) Return only articles in the following subject areas:
  - [ ] Biology, Life Sciences, and Environmental Science
  - [ ] Business, Administration, Finance, and Economics
  - [ ] Chemistry and Materials Science
  - [ ] Engineering, Computer Science, and Mathematics
  - [ ] Medicine, Pharmacology, and Veterinary Science
  - [ ] Physics, Astronomy, and Planetary Science
  - [ ] Social Sciences, Arts, and Humanities

Google Scholar BETA

Advanced Scholar Search

Advanced Search Tips | About Google Scholar

Find articles with **all** of the words

with the **exact phrase**

with **at least one** of the words

without the words

where my words occur: anywhere in the article

10 results | Search Scholar

Author Return articles written by: abiteboul

e.g., *"PJ Hayes"* or *McCarthy*

Publication Return articles published in

e.g., *J Biol Chem* or *Nature*

Date Return articles published between —

e.g., *1996*

Subject Areas

- (●) Return articles in all subject areas.
- ( ) Return only articles in the following subject areas:
  - [ ] Biology, Life Sciences, and Environmental Science
  - [ ] Business, Administration, Finance, and Economics
  - [ ] Chemistry and Materials Science
  - [ ] Engineering, Computer Science, and Mathematics
  - [ ] Medicine, Pharmacology, and Veterinary Science
  - [ ] Physics, Astronomy, and Planetary Science
  - [ ] Social Sciences, Arts, and Humanities

Google Scholar BETA

author:abiteboul Search

Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar **All articles - Recent articles** Results **21** - **30** of about **950** for **author:abiteboul**. (**0.13** seconds)

abiteboul
S Abiteboul
V Vianu
J McHugh
R Hull
J Widom

Querying documents in object databases - all 3 versions »
S **Abiteboul**, S Cluet, V Christophides, T Milo, G … - International Journal on Digital Libraries, 1997 - Springer
Abstract. We consider the problem of storing and accessing documents (SGML and HTML, in particular) using database technology. To specify the database image of documents, we use structuring schemas that consist in grammars annotated ...
Cited by 159 - Related Articles - Web Search

Generic Computation and its complexity - all 2 versions »
S **Abiteboul**, V Vianu - Proceedings of the twenty-third annual ACM symposium on …, 1991 - portal.acm.org
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, ...
Cited by 156 - Related Articles - Web Search

[PDF] Change-Centric Management of Versions in an XML Warehouse - all 10 versions »
A Marian, S **Abiteboul**, G Cobena, L Mignet - Proceedings of VLDB 2001, 2001 - gregory.cobena.free.fr
Abstract: We present a change-centric method to manage versions in a Web WareHouse of XML data. The starting points is a sequence of snapshots of XML documents we obtain from the web. By running a diff algorithm, we compute ...
Cited by 154 - Related Articles - View as HTML - Web Search

[BOOK] COL: A Logic-Based Language for Complex Objects - all 6 versions »
S **Abiteboul**, S Grumbach - Springer
Abstract: A logic-based language for manipalating complex objects constructed using set and tuple constructors is introduced. A key feature of the language is the use of base and derived data functions. Under some stratification res- ...
Cited by 150 - Related Articles - Web Search - SUDOC Catalogue

Google Scholar BETA

author:abiteboul Search

Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar **All articles - Recent articles** Results **21** - **30** of about **950** for **author:abiteboul**. (**0.13** seconds)

abiteboul
S Abiteboul
V Vianu
J McHugh
R Hull
J Widom

Querying documents in object databases - all 3 versions »
S **Abiteboul**, S Cluet, V Christophides, T Milo, G … - International Journal on Digital Libraries, 1997 - Springer
Abstract. We consider the problem of storing and accessing documents (SGML and HTML, in particular) using database technology. To specify the database image of documents, we use structuring schemas that consist in grammars annotated ...
Cited by 159 - Related Articles - Web Search

Generic Computation and its complexity - all 2 versions »
S **Abiteboul**, V Vianu - Proceedings of the twenty-third annual ACM symposium on …, 1991 - portal.acm.org
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, ...
Cited by 156 - Related Articles - Web Search

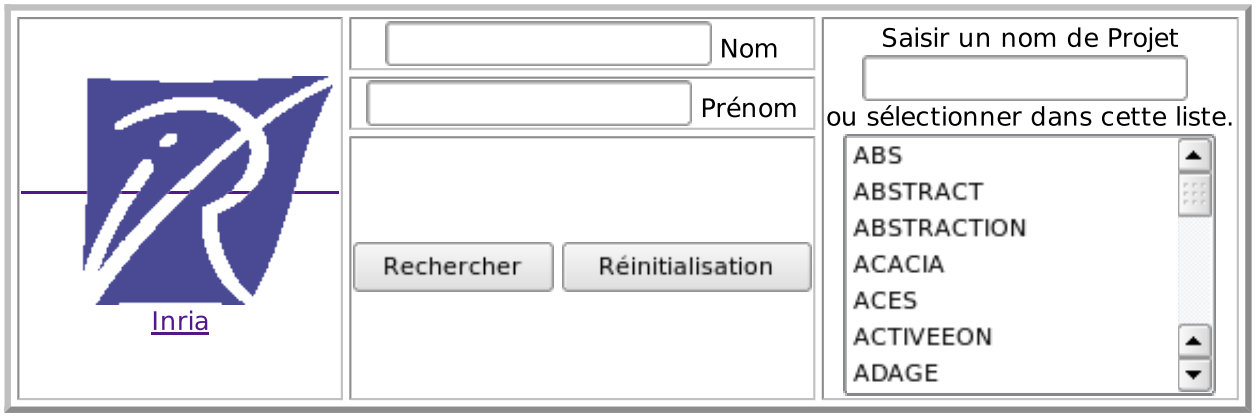
[PDF] Change-Centric Management of Versions in an XML Warehouse - all 10 versions »
A Marian, S **Abiteboul**, G Cobena, L Mignet - Proceedings of VLDB 2001, 2001 - gregory.cobena.free.fr
Abstract: We present a change-centric method to manage versions in a Web WareHouse of XML data. The starting points is a sequence of snapshots of XML documents we obtain from the web. By running a diff algorithm, we compute ...
Cited by 154 - Related Articles - View as HTML - Web Search

[BOOK] COL: A Logic-Based Language for Complex Objects - all 6 versions »
S **Abiteboul**, S Grumbach - Springer
Abstract: A logic-based language for manipalating complex objects constructed using set and tuple constructors is introduced. A key feature of the language is the use of base and derived data functions. Under some stratification res- ...
Cited by 150 - Related Articles - Web Search - SUDOC Catalogue

Annuaire de l'Inria.

Annuaire (prototype) des personnes travaillant à l'Inria. Les données de cet annuaire sont issues des bases des autocommutateurs de l'Inria.

On peut saisir un ou plusieurs champs. Pour élargir votre requête, préfixez la par une *

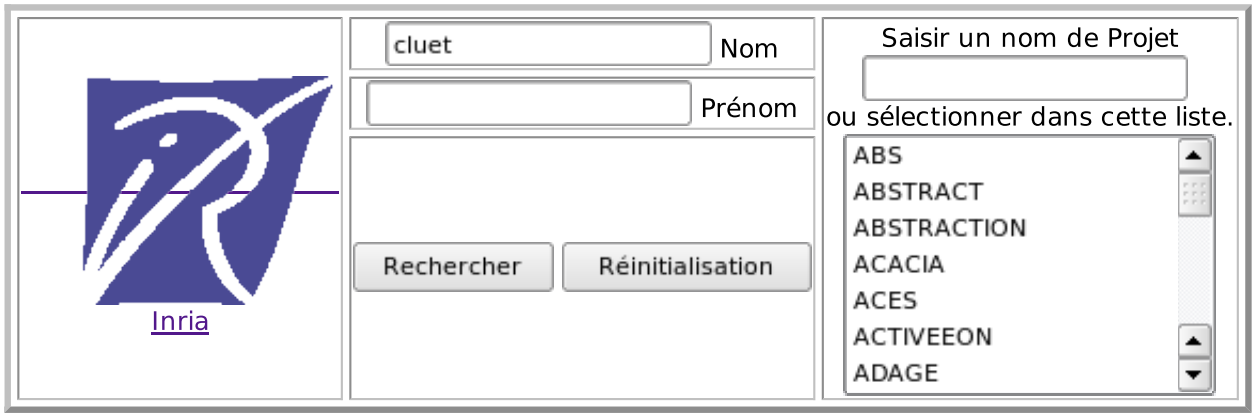


Poster vos questions et commentaires,
sauf les demandes de mises à jour qui doivent être effectuées auprès du service informatique de votre site.

Annuaire de l'Inria.

Annuaire (prototype) des personnes travaillant à l'Inria. Les données de cet annuaire sont issues des bases des autocommutateurs de l'Inria.

On peut saisir un ou plusieurs champs. Pour élargir votre requête, préfixez la par une *



Poster vos questions et commentaires,
sauf les demandes de mises à jour qui doivent être effectuées auprès du service informatique de votre site.

Annuaire de l'Inria

1 personne(s) trouvée(s).

Prénom Nom	Courrier électronique	Téléphone	Site	Projet
Sophie Cluet	Sophie.Cluet@inria.fr		futurs	GEMO

Interroger à nouveau l'annuaire Inria ***Poster vos questions et commentaires*** .

Annuaire de l'Inria

1 personne(s) trouvée(s).

Prénom Nom	Courrier électronique	Téléphone	Site	Projet
Sophie Cluet	Sophie.Cluet@inria.fr		futurs	GEMO

Interroger à nouveau l'annuaire Inria ***Poster vos questions et commentaires*** .

Google Scholar BETA

author:abiteboul Search

Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar **All articles - Recent articles** Results **21** - **30** of about **950** for **author:abiteboul**. (**0.13** seconds)

abiteboul
S Abiteboul
V Vianu
J McHugh
R Hull
J Widom

Querying documents in object databases - all 3 versions »
S **Abiteboul**, S Cluet, V Christophides, T Milo, G … - International Journal on Digital Libraries, 1997 - Springer
Abstract. We consider the problem of storing and accessing documents (SGML and HTML, in particular) using database technology. To specify the database image of documents, we use structuring schemas that consist in grammars annotated ...
Cited by 159 - Related Articles - Web Search

Generic Computation and its complexity - all 2 versions »
S **Abiteboul**, V Vianu - Proceedings of the twenty-third annual ACM symposium on …, 1991 - portal.acm.org
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, ...
Cited by 156 - Related Articles - Web Search

[PDF] Change-Centric Management of Versions in an XML Warehouse - all 10 versions »
A Marian, S **Abiteboul**, G Cobena, L Mignet - Proceedings of VLDB 2001, 2001 - gregory.cobena.free.fr
Abstract: We present a change-centric method to manage versions in a Web WareHouse of XML data. The starting points is a sequence of snapshots of XML documents we obtain from the web. By running a diff algorithm, we compute ...
Cited by 154 - Related Articles - View as HTML - Web Search

[BOOK] COL: A Logic-Based Language for Complex Objects - all 6 versions »
S **Abiteboul**, S Grumbach - Springer
Abstract: A logic-based language for manipalating complex objects constructed using set and tuple constructors is introduced. A key feature of the language is the use of base and derived data functions. Under some stratification res- ...
Cited by 150 - Related Articles - Web Search - SUDOC Catalogue

Google Scholar BETA

author:abiteboul Search

Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar **All articles - Recent articles** Results **21** - **30** of about **950** for **author:abiteboul**. (**0.13** seconds)

abiteboul
S Abiteboul
V Vianu
J McHugh
R Hull
J Widom

Querying documents in object databases - all 3 versions »
S **Abiteboul**, S Cluet, V Christophides, T Milo, G … - International Journal on Digital Libraries, 1997 - Springer
Abstract. We consider the problem of storing and accessing documents (SGML and HTML, in particular) using database technology. To specify the database image of documents, we use structuring schemas that consist in grammars annotated ...
Cited by 159 - Related Articles - Web Search

Generic Computation and its complexity - all 2 versions »
S **Abiteboul**, V Vianu - Proceedings of the twenty-third annual ACM symposium on …, 1991 - portal.acm.org
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, ...
Cited by 156 - Related Articles - Web Search

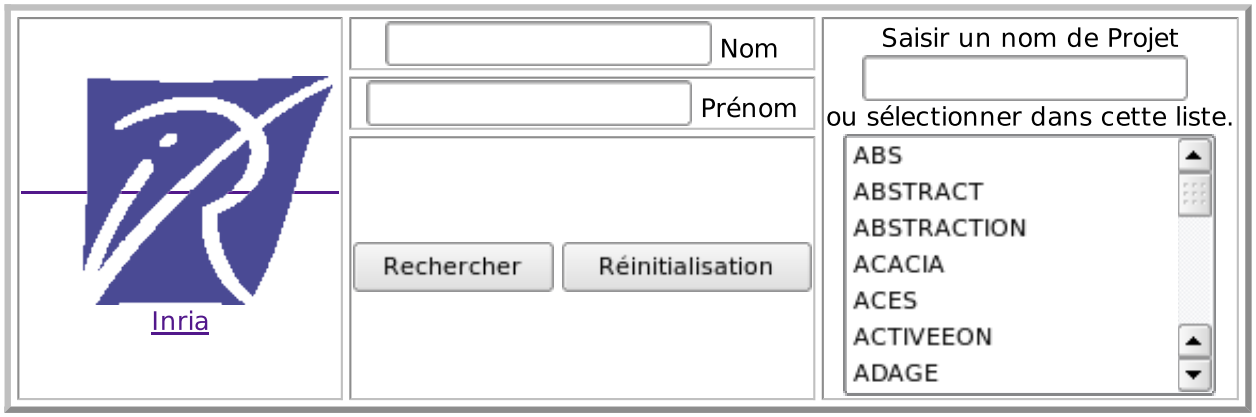
[PDF] Change-Centric Management of Versions in an XML Warehouse - all 10 versions »
A Marian, S **Abiteboul**, G Cobena, L Mignet - Proceedings of VLDB 2001, 2001 - gregory.cobena.free.fr
Abstract: We present a change-centric method to manage versions in a Web WareHouse of XML data. The starting points is a sequence of snapshots of XML documents we obtain from the web. By running a diff algorithm, we compute ...
Cited by 154 - Related Articles - View as HTML - Web Search

[BOOK] COL: A Logic-Based Language for Complex Objects - all 6 versions »
S **Abiteboul**, S Grumbach - Springer
Abstract: A logic-based language for manipalating complex objects constructed using set and tuple constructors is introduced. A key feature of the language is the use of base and derived data functions. Under some stratification res- ...
Cited by 150 - Related Articles - Web Search - SUDOC Catalogue

Annuaire de l'Inria.

Annuaire (prototype) des personnes travaillant à l'Inria. Les données de cet annuaire sont issues des bases des autocommutateurs de l'Inria.

On peut saisir un ou plusieurs champs. Pour élargir votre requête, préfixez la par une *

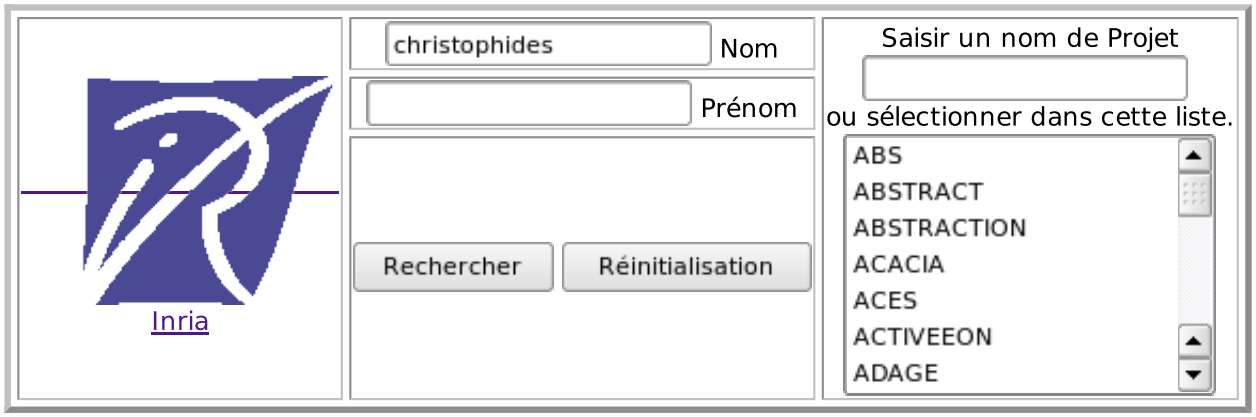


Poster vos questions et commentaires,
sauf les demandes de mises à jour qui doivent être effectuées auprès du service informatique de votre site.

Annuaire de l'Inria.

Annuaire (prototype) des personnes travaillant à l'Inria. Les données de cet annuaire sont issues des bases des autocommutateurs de l'Inria.

On peut saisir un ou plusieurs champs. Pour élargir votre requête, préfixez la par une *



Poster vos questions et commentaires,
sauf les demandes de mises à jour qui doivent être effectuées auprès du service informatique de votre site.

Annuaire de l'Inria

0 personne(s) trouvée(s).

Prénom Nom Courrier électronique Téléphone Site Projet

Interroger à nouveau l'annuaire Inria ***Poster vos questions et commentaires*** .

Lessons learned

Tedious!

- Identify all relevant services.
- Understand the way to query them.
- Query them one by one.
- Merge the results.

Goal

Have all this work done by the computer, in a fully automatic way.

The hidden Web

Définition (Hidden Web, Deep Web, Invisible Web)

All contents of the Web that are not accessible through the hyperlinked structure of the World Wide Web. Generally speaking: HTML forms, Web services.

Size estimate (2001): 500 more data than on the surface Web.

How to understand and benefit from this content?

Understanding the hidden Web

Goal

- Intensional indexing of the hidden Web.
- High-level queries.
- ⇒ a semantic search engine for the hidden Web.

In a fully automatic
and unsupervised way!

- Difficult and broad problem.
- Use of domain knowledge (ontology, instances).
- Example with the domain of research publications.

Understanding the hidden Web

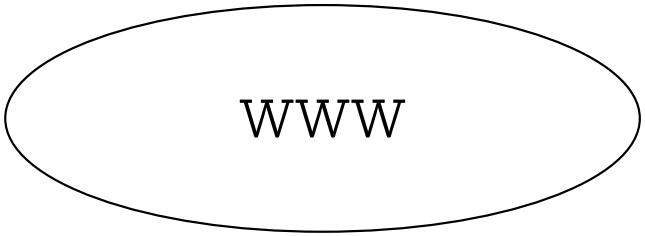
Goal

- Intensional indexing of the hidden Web.
- High-level queries.
- ⇒ a semantic search engine for the hidden Web.

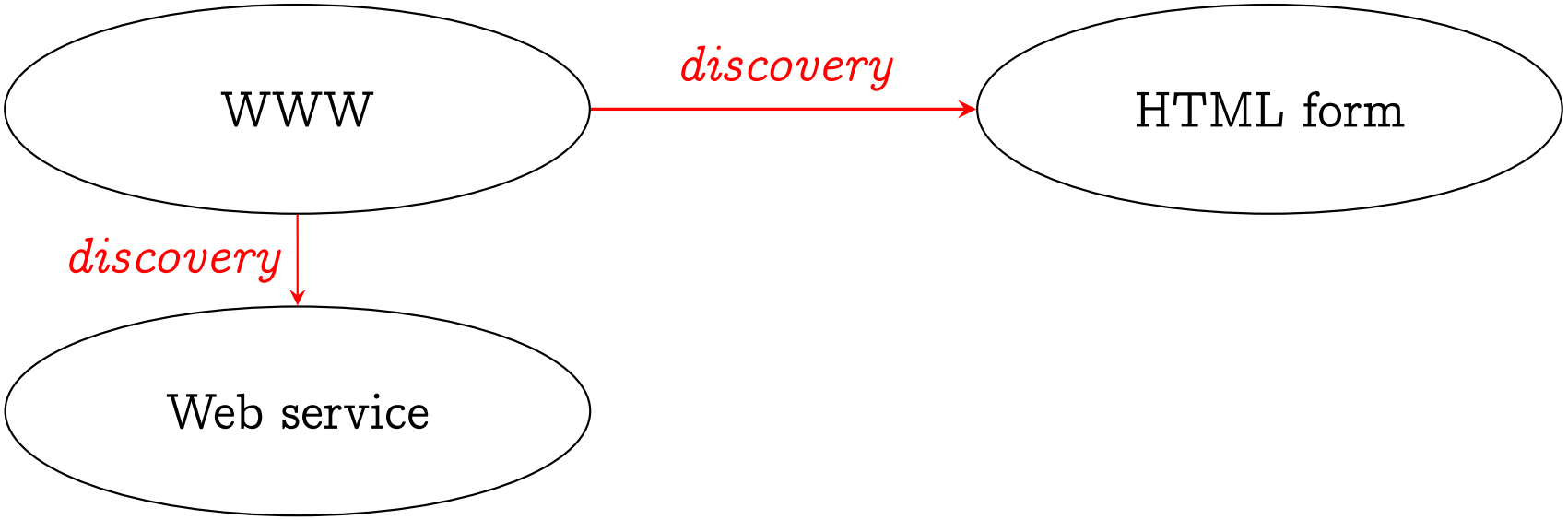
In a fully automatic
and unsupervised way!

- Difficult and broad problem.
- Use of domain knowledge (ontology, instances).
- Example with the domain of research publications.

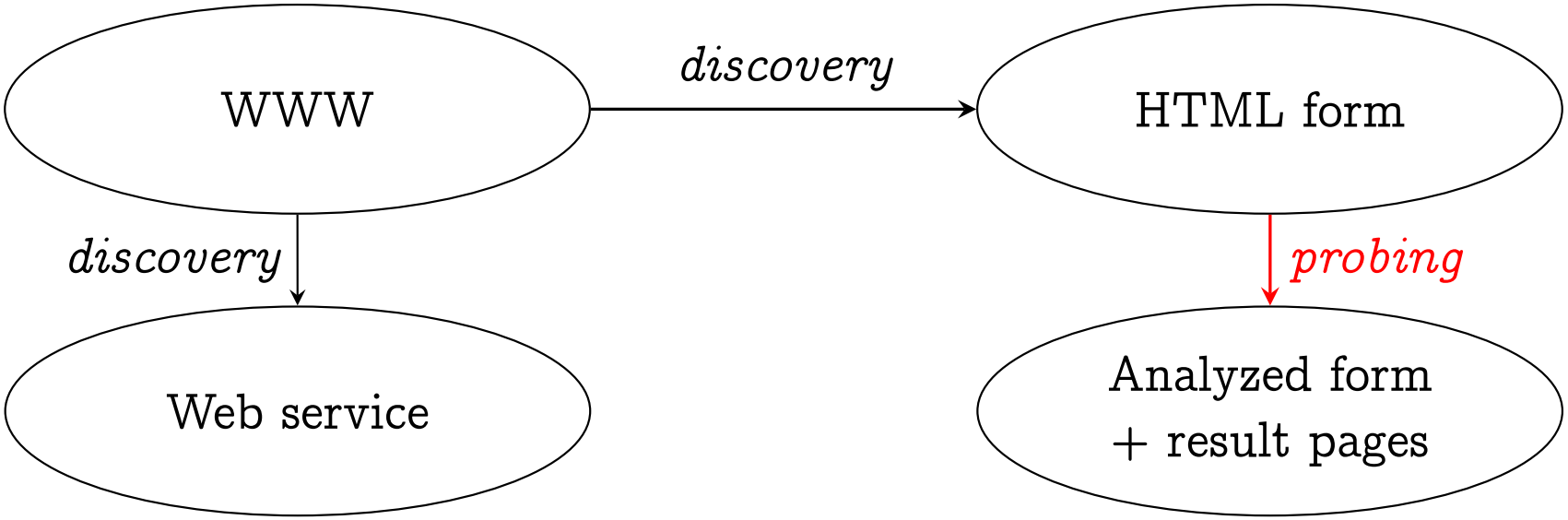
Process for the semantic interpretation of the hidden Web



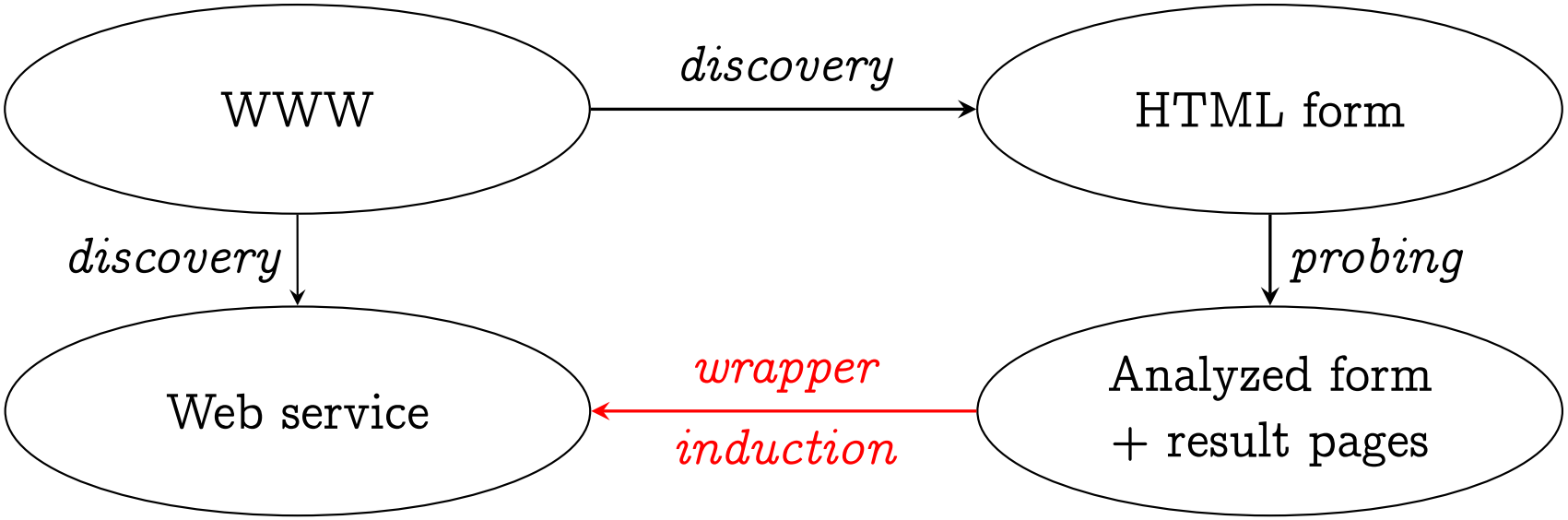
Process for the semantic interpretation of the hidden Web



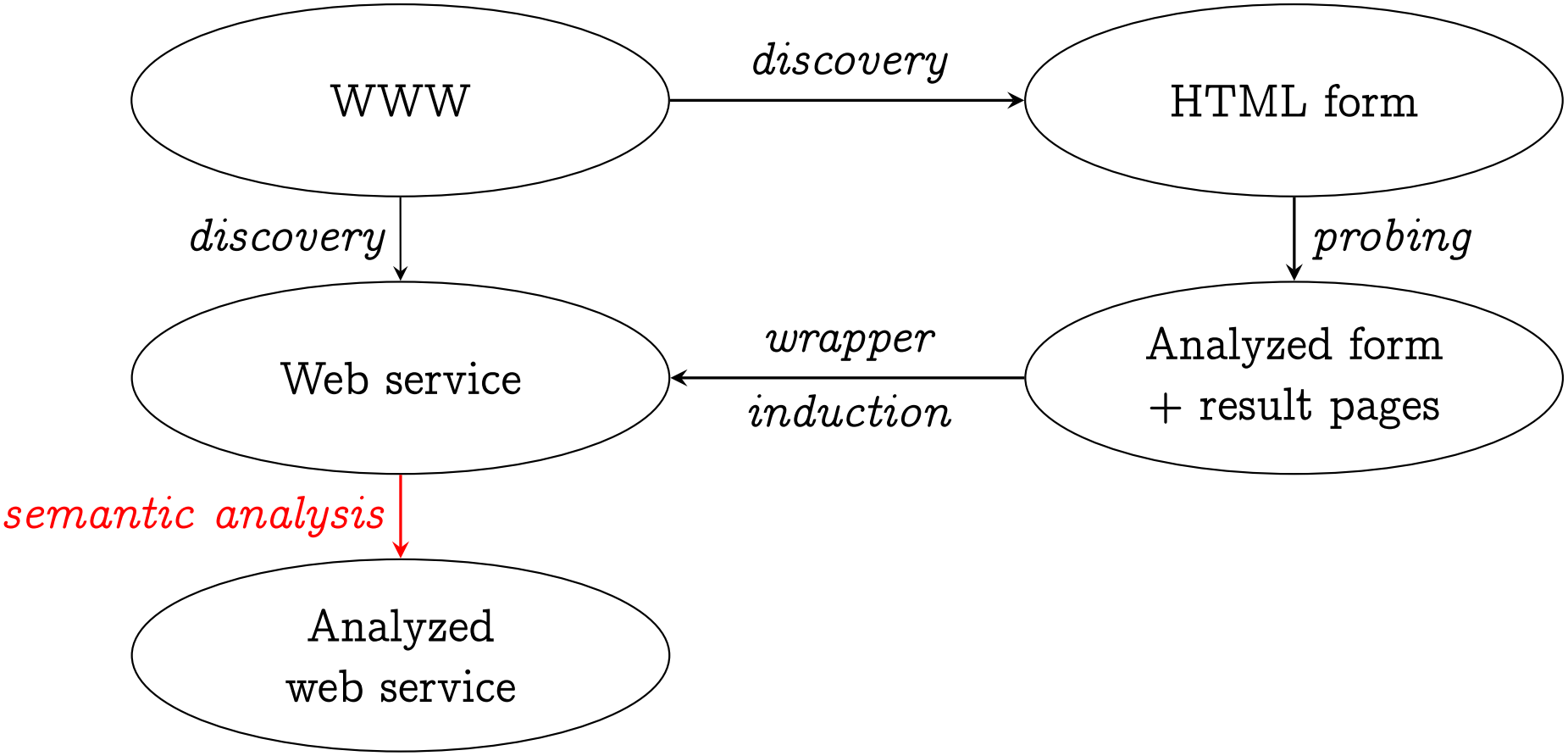
Process for the semantic interpretation of the hidden Web



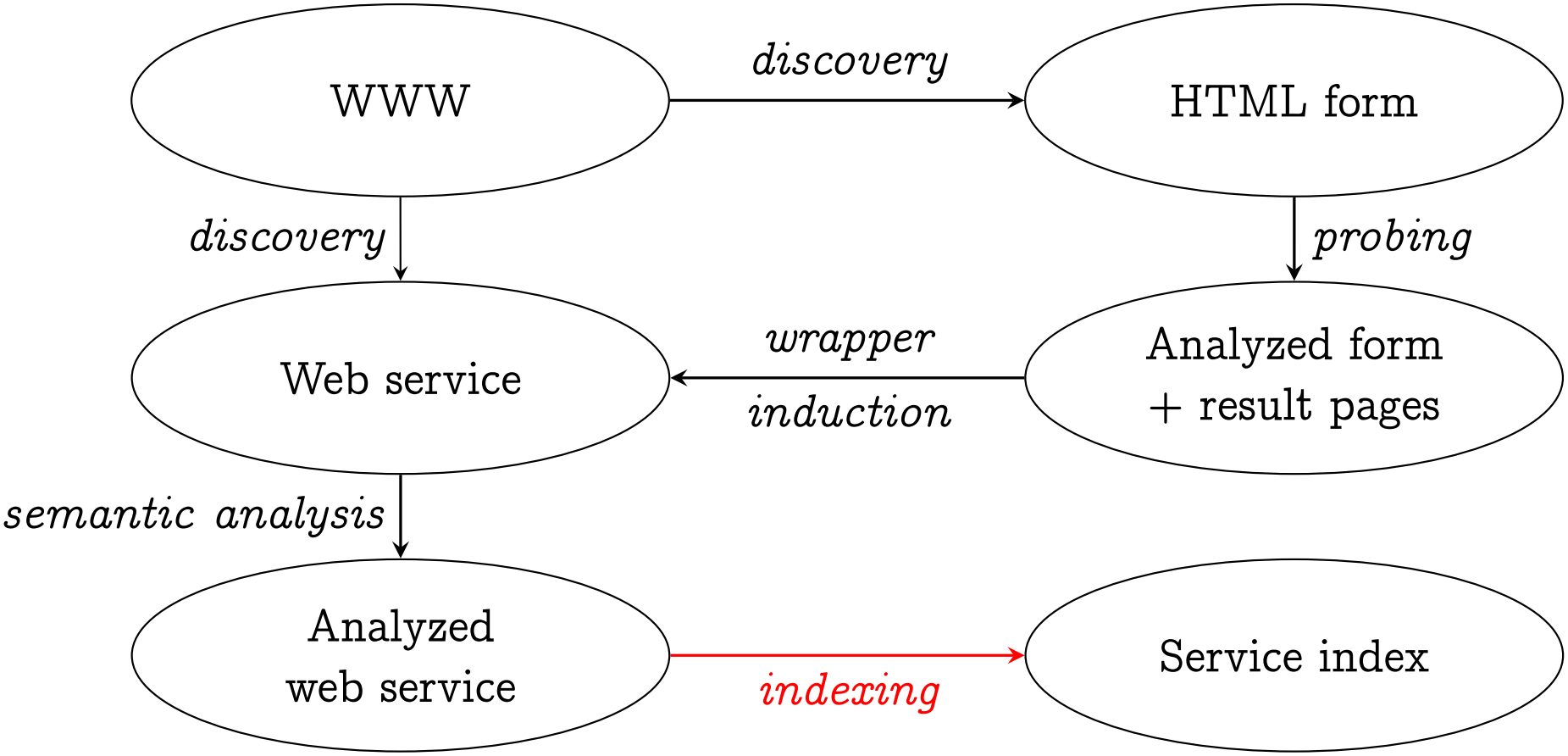
Process for the semantic interpretation of the hidden Web



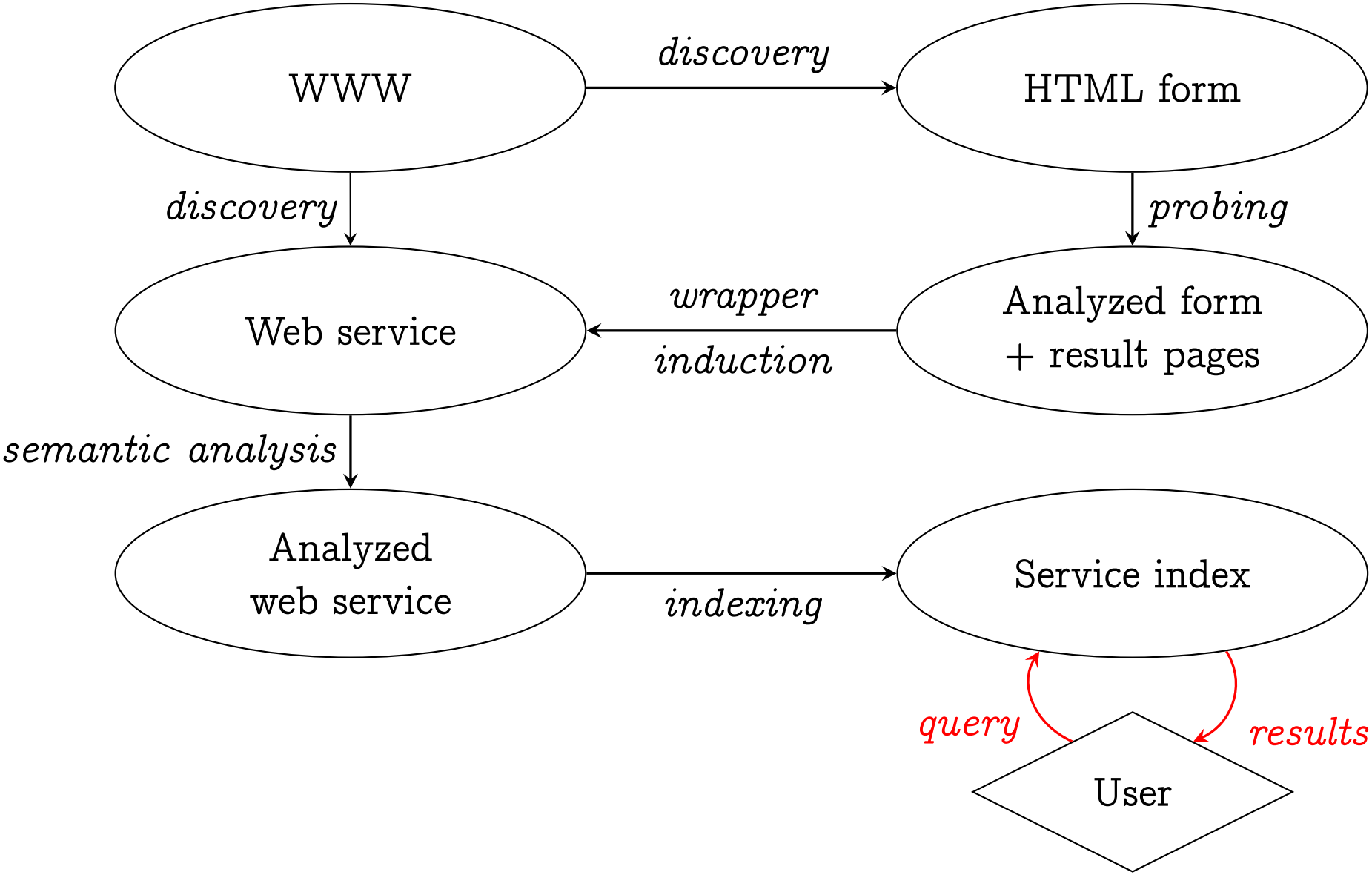
Process for the semantic interpretation of the hidden Web



Process for the semantic interpretation of the hidden Web



Process for the semantic interpretation of the hidden Web

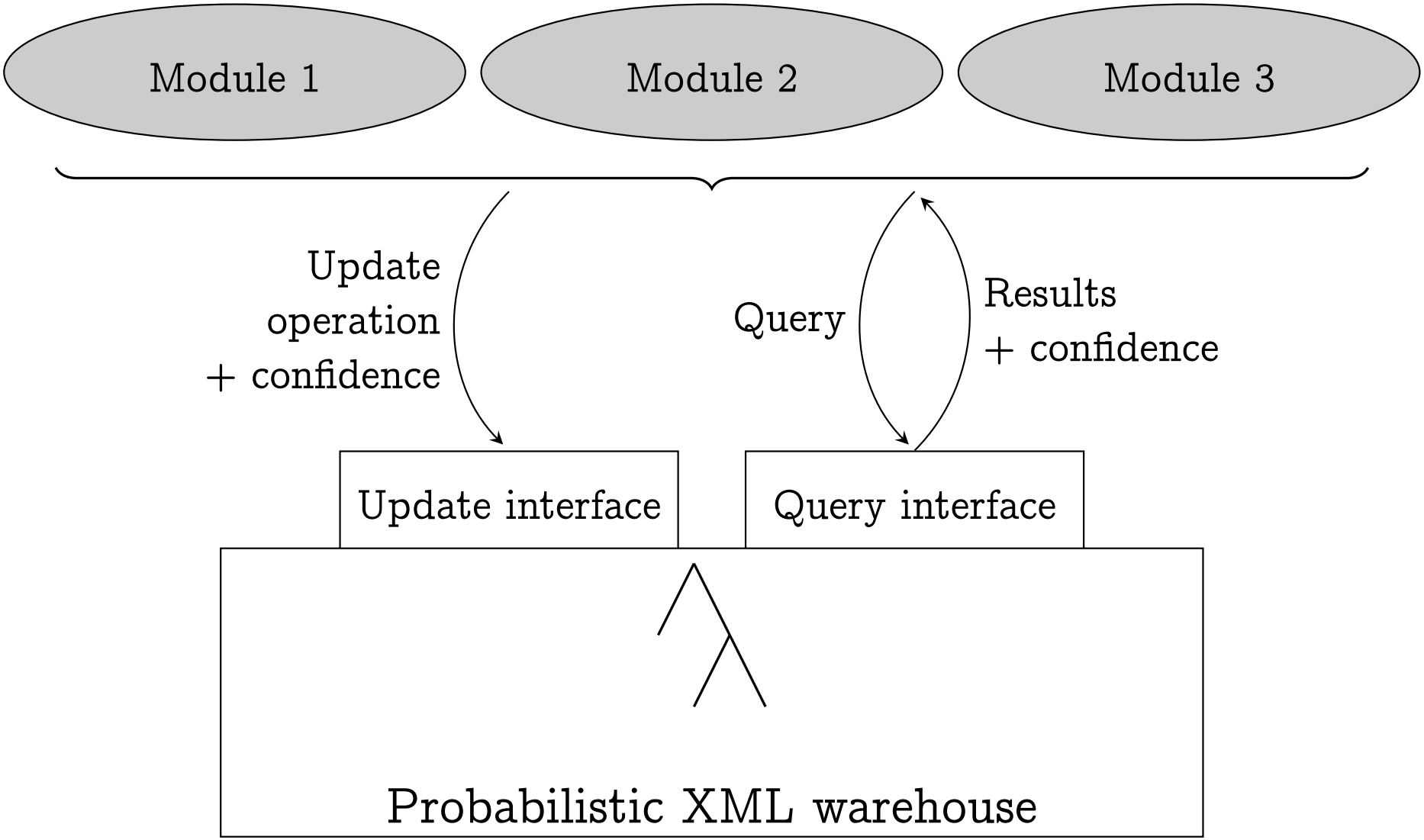


Imprecise data and tasks

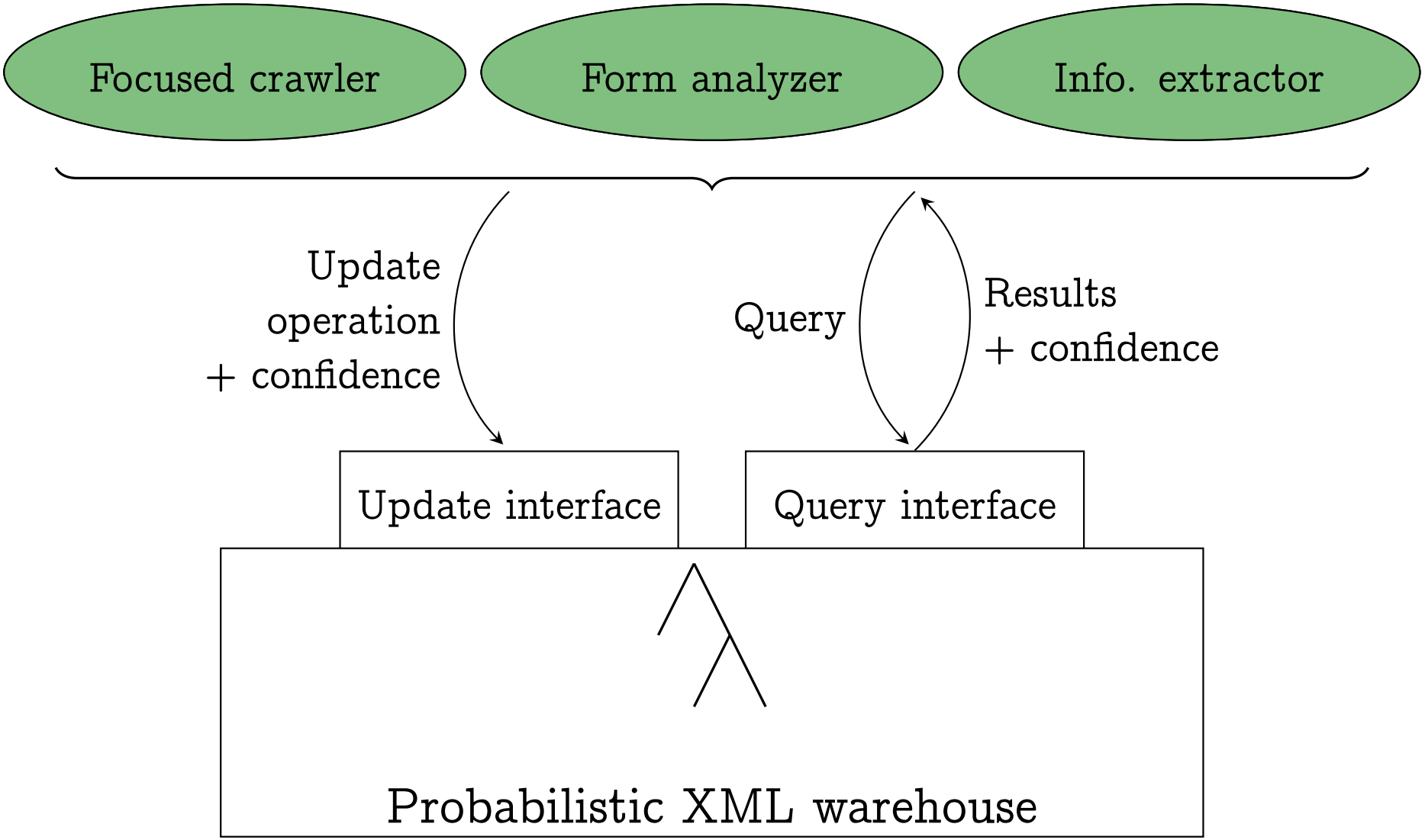
Observations

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

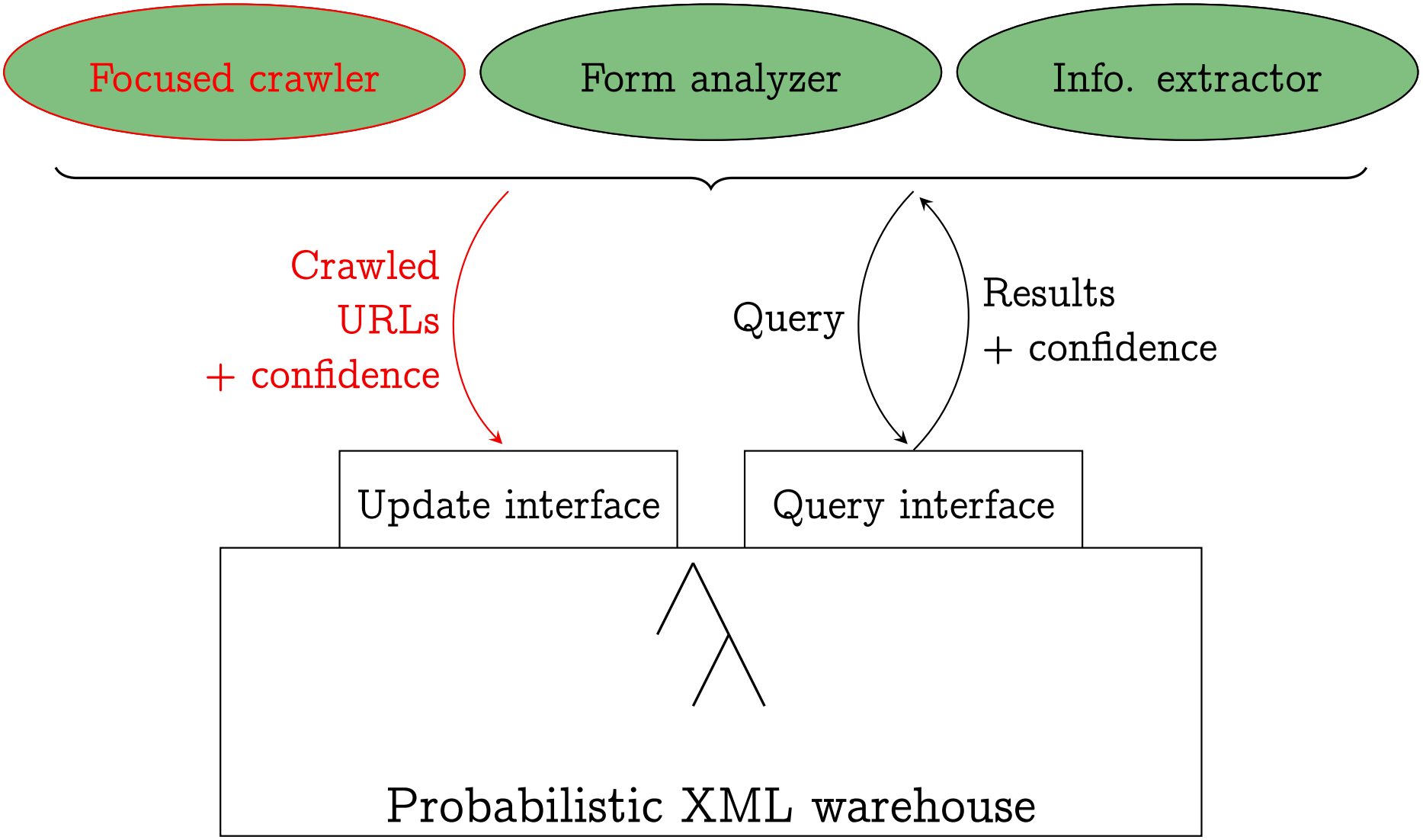
A probabilistic XML warehouse



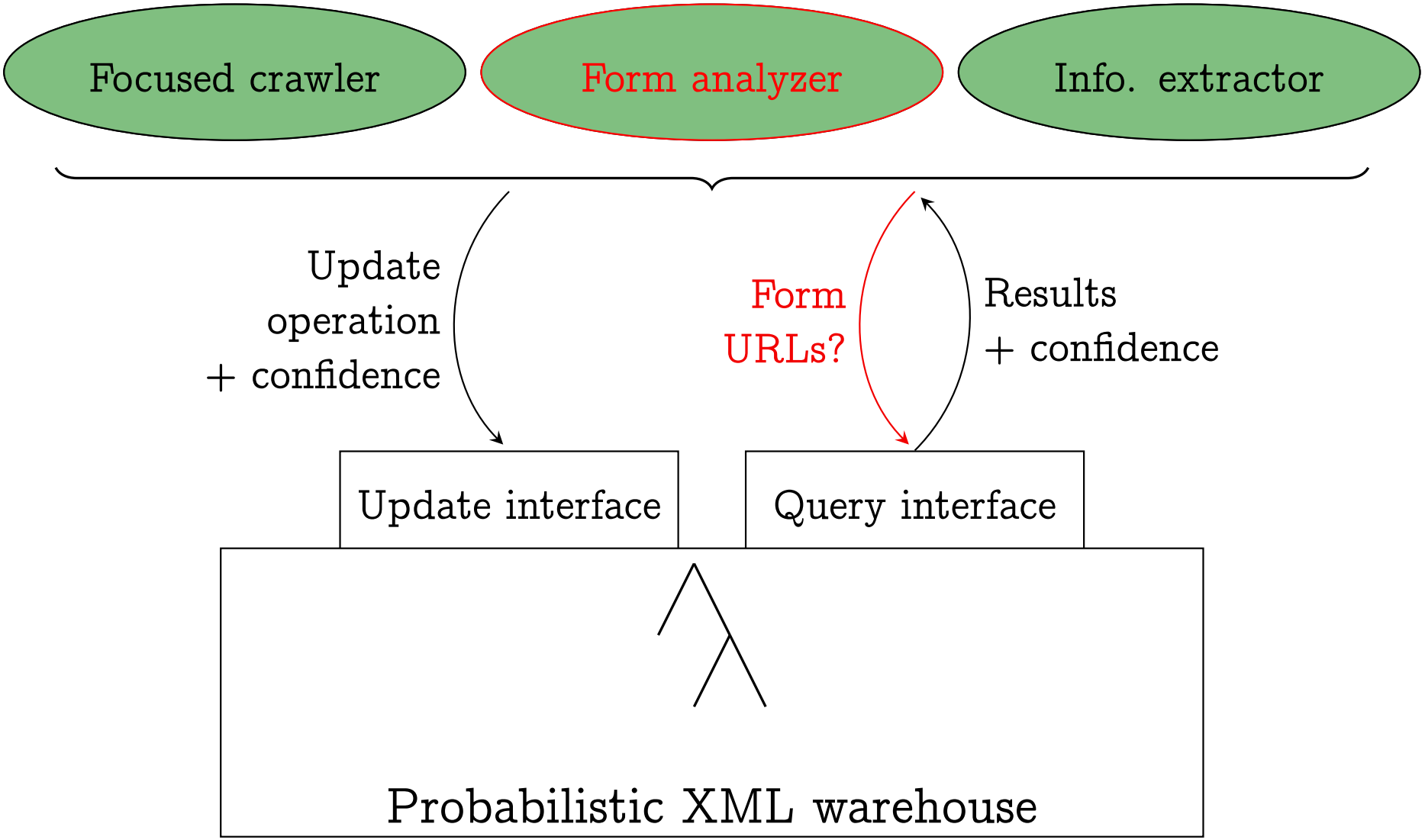
A probabilistic XML warehouse (hidden Web)



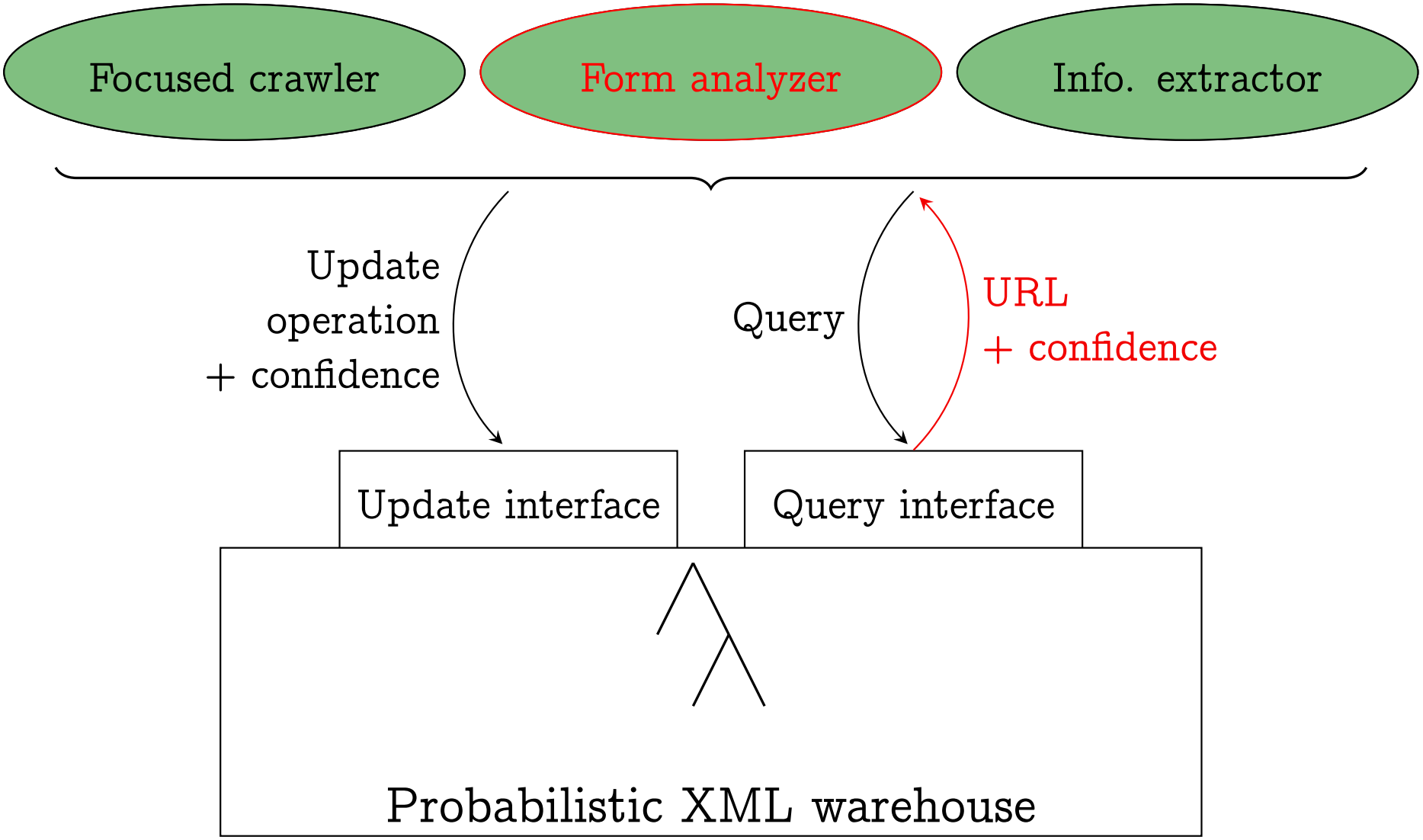
A probabilistic XML warehouse (hidden Web)



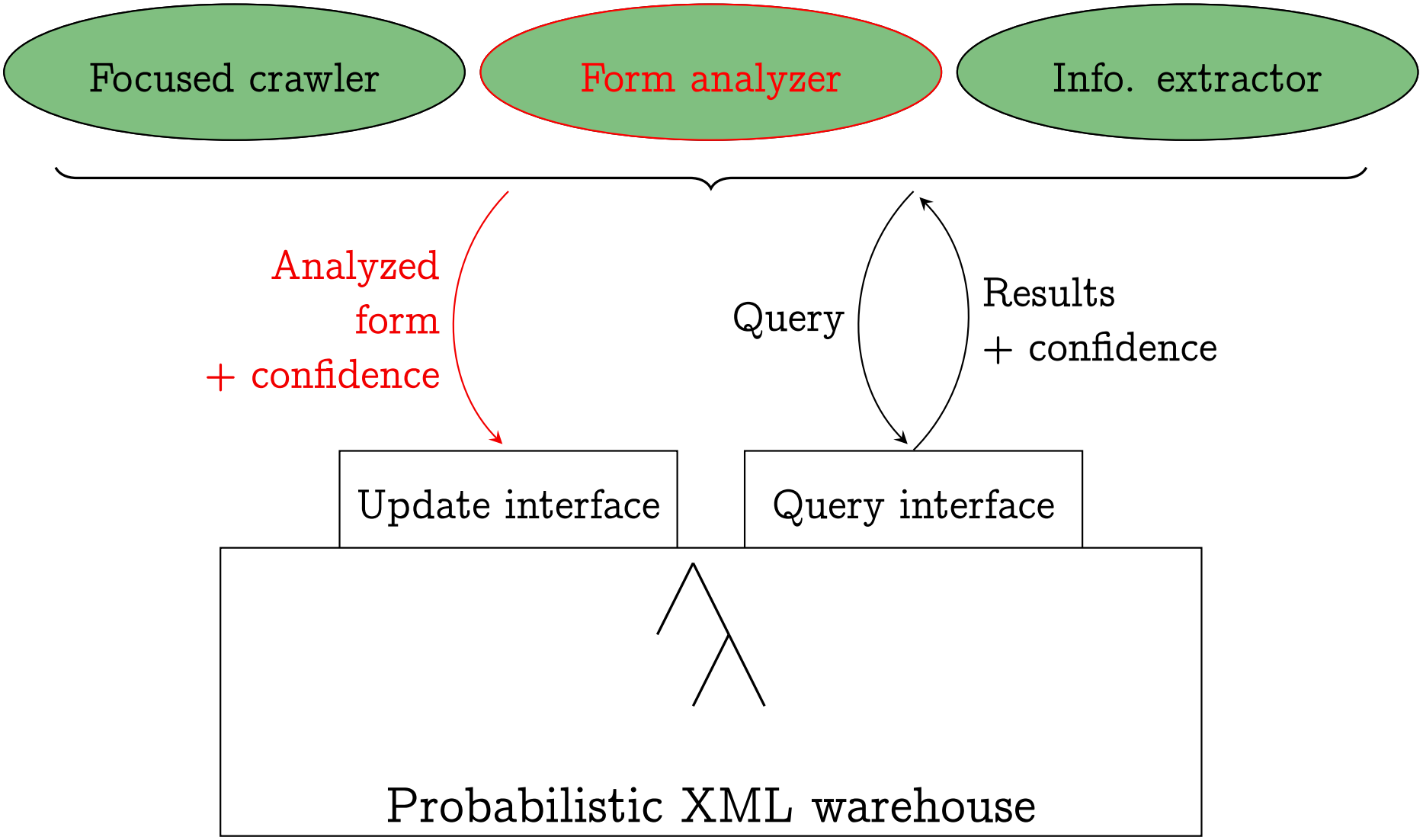
A probabilistic XML warehouse (hidden Web)



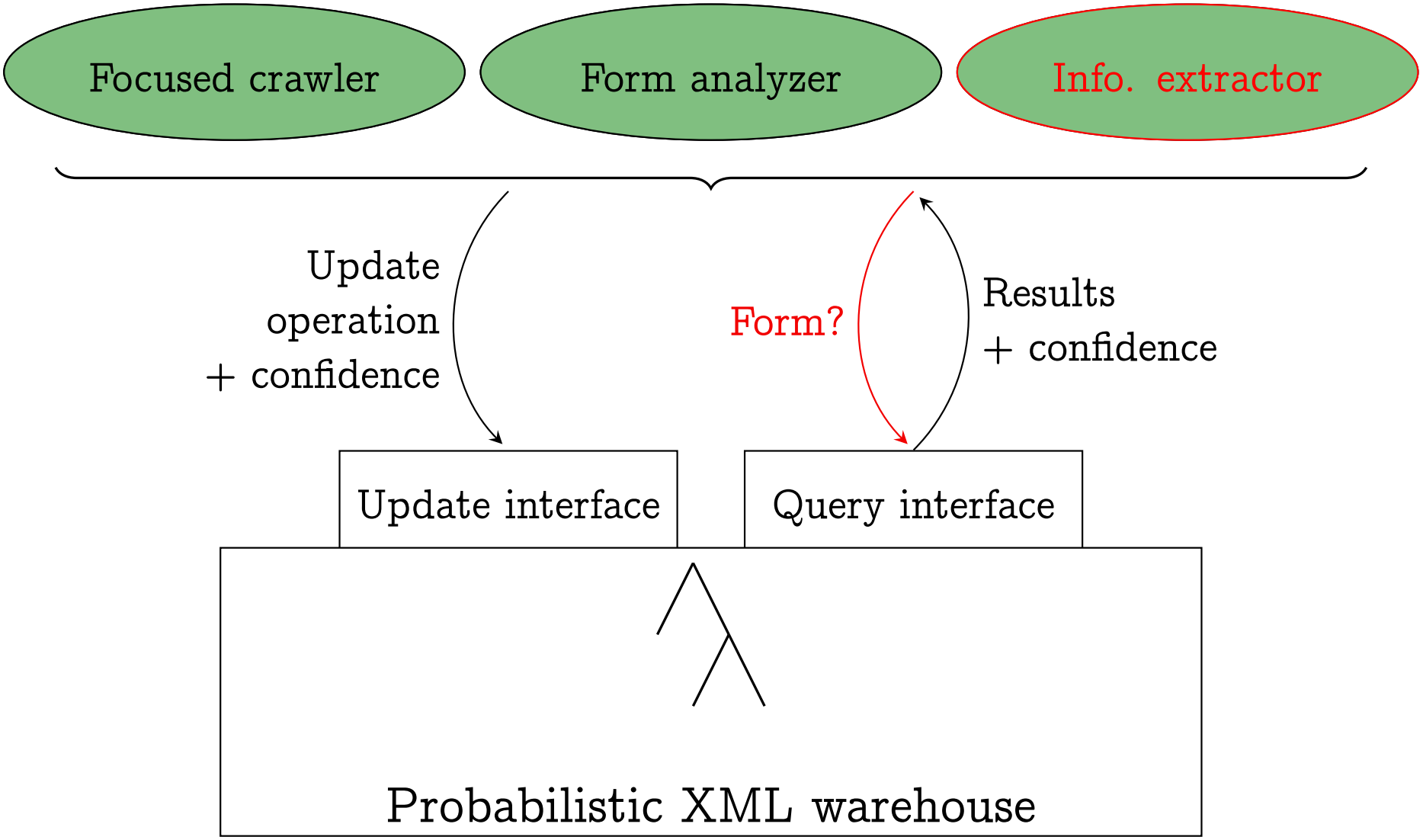
A probabilistic XML warehouse (hidden Web)



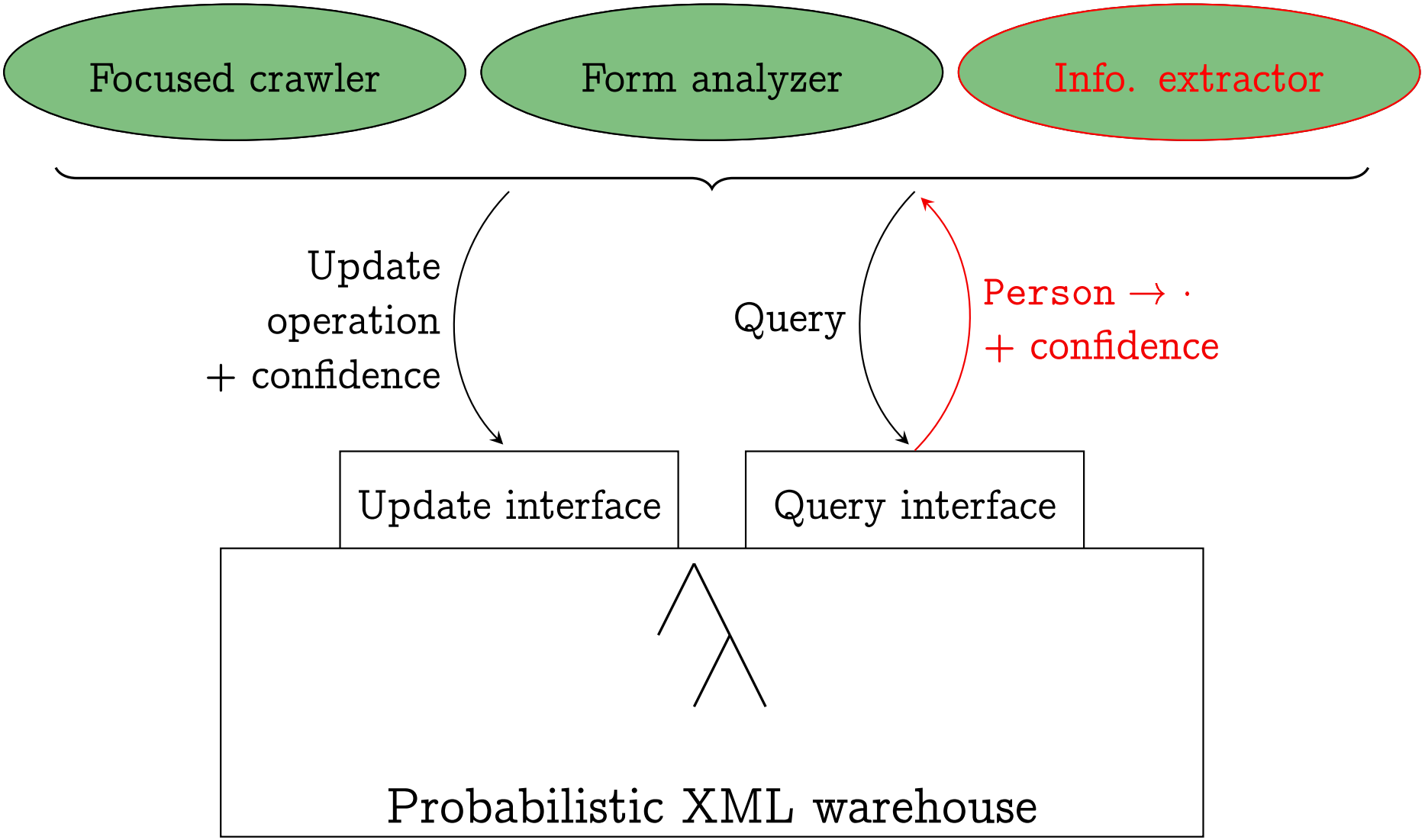
A probabilistic XML warehouse (hidden Web)



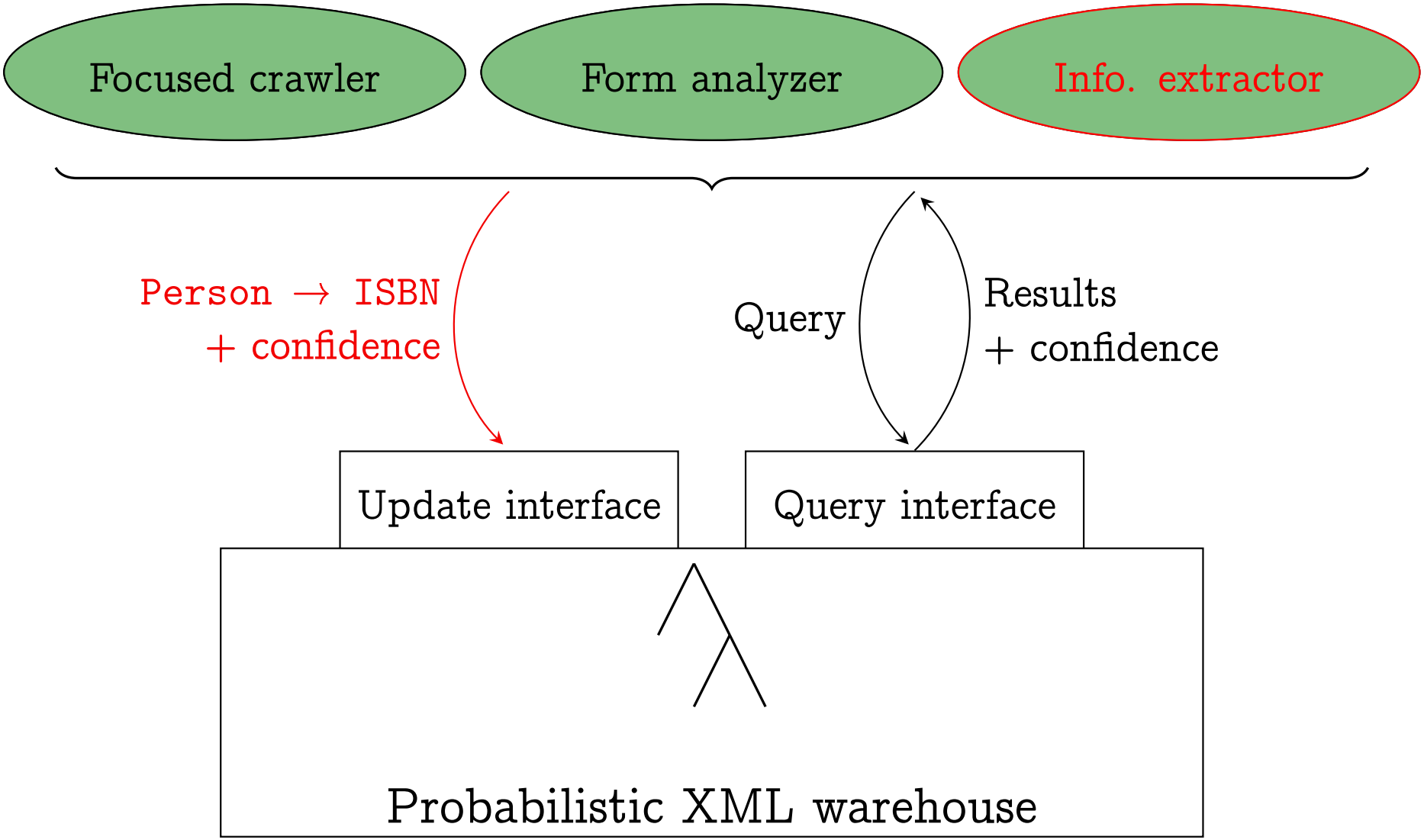
A probabilistic XML warehouse (hidden Web)



A probabilistic XML warehouse (hidden Web)



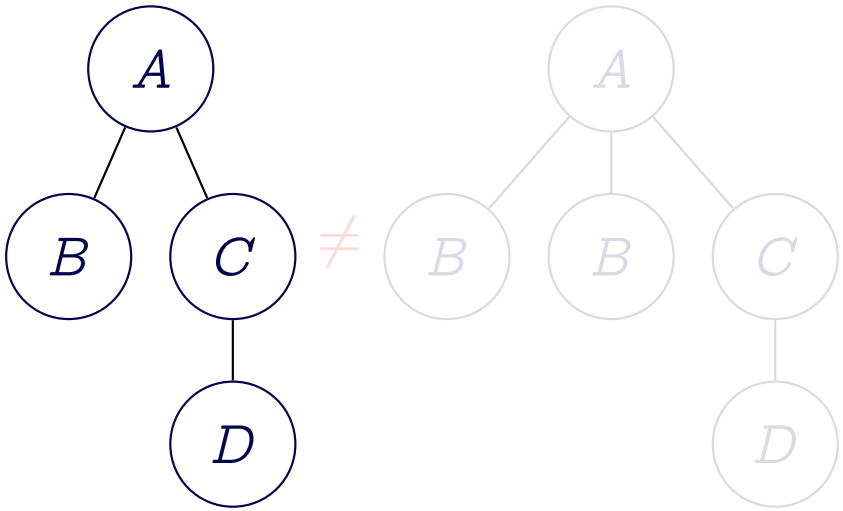
A probabilistic XML warehouse (hidden Web)



Probabilistic trees

Framework

- Unordered data trees.
- Simplifications (without loss of generality): no attributes, no mixed content...



(multiset semantics)

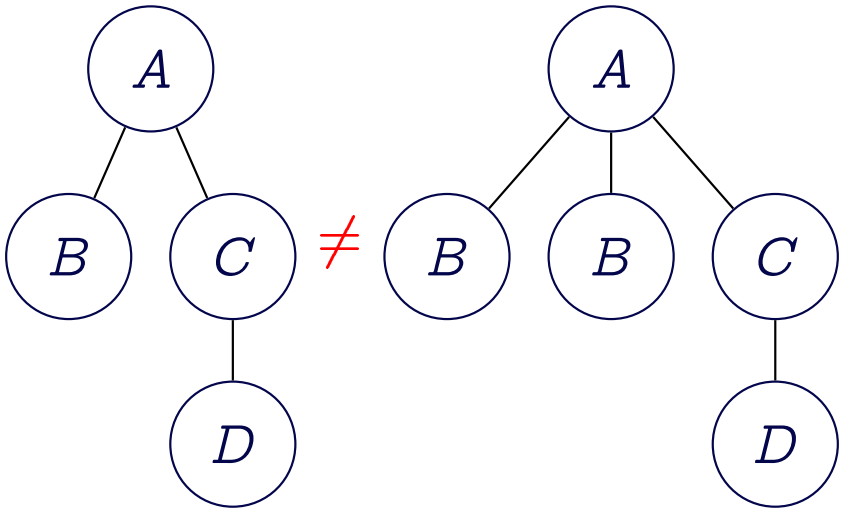
Universe: Set of all such data trees.

Probabilistic tree: Syntactic representation of a discrete probability distribution in this universe.

Probabilistic trees

Framework

- Unordered data trees.
- Simplifications (without loss of generality): no attributes, no mixed content...



(multiset semantics)

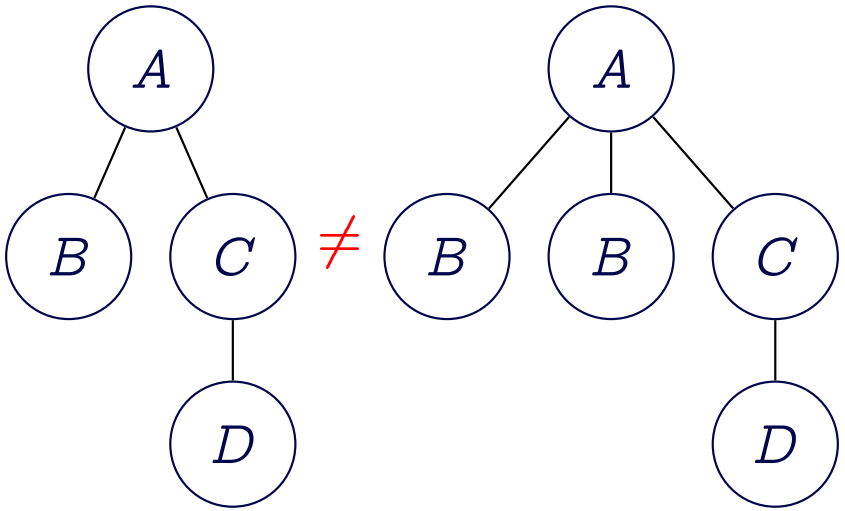
Universe: Set of all such data trees.

Probabilistic tree: Syntactic representation of a discrete probability distribution in this universe.

Probabilistic trees

Framework

- Unordered data trees.
- Simplifications (without loss of generality): no attributes, no mixed content...



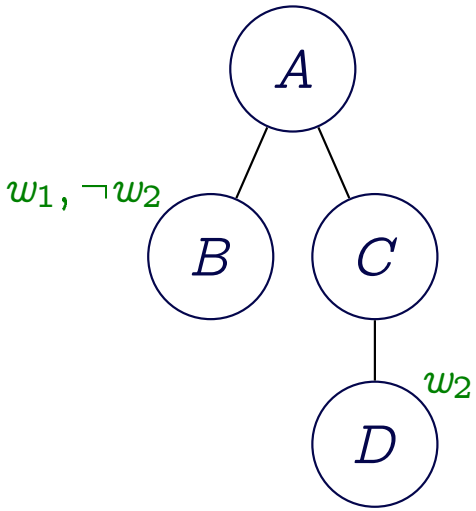
(multiset semantics)

Universe: Set of all such data trees.

Probabilistic tree: Syntactic representation of a discrete probability distribution in this universe.

Probabilistic tree model

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

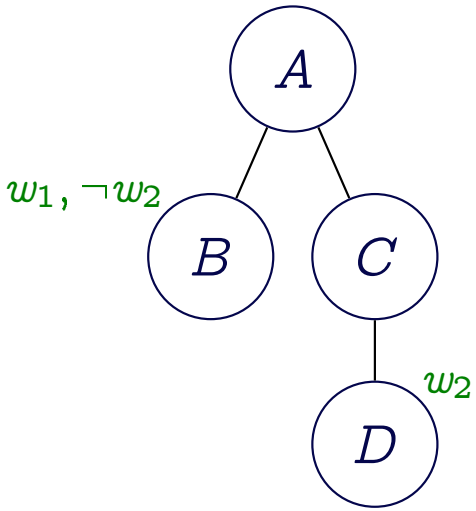


Event	Prob.
w_1	0.8
w_2	0.7

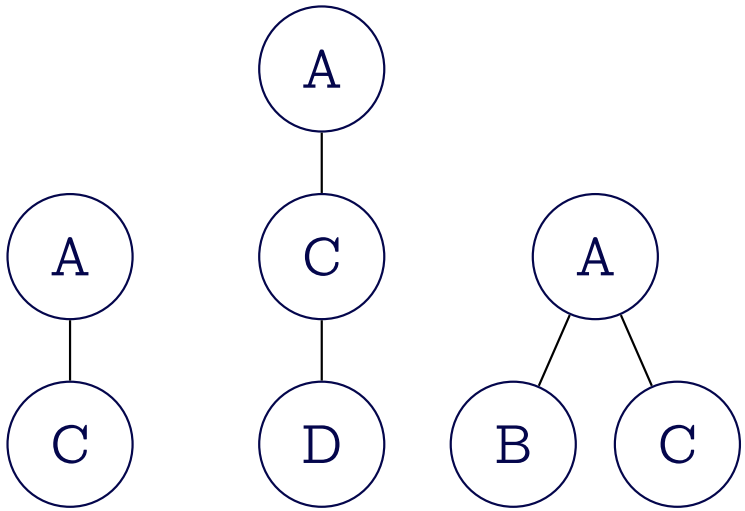
semantics →

Probabilistic tree model

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



semantics →



$p_1 = 0.06$ $p_2 = 0.70$ $p_3 = 0.24$

Event	Prob.
w_1	0.8
w_2	0.7

Features of the probabilistic tree model

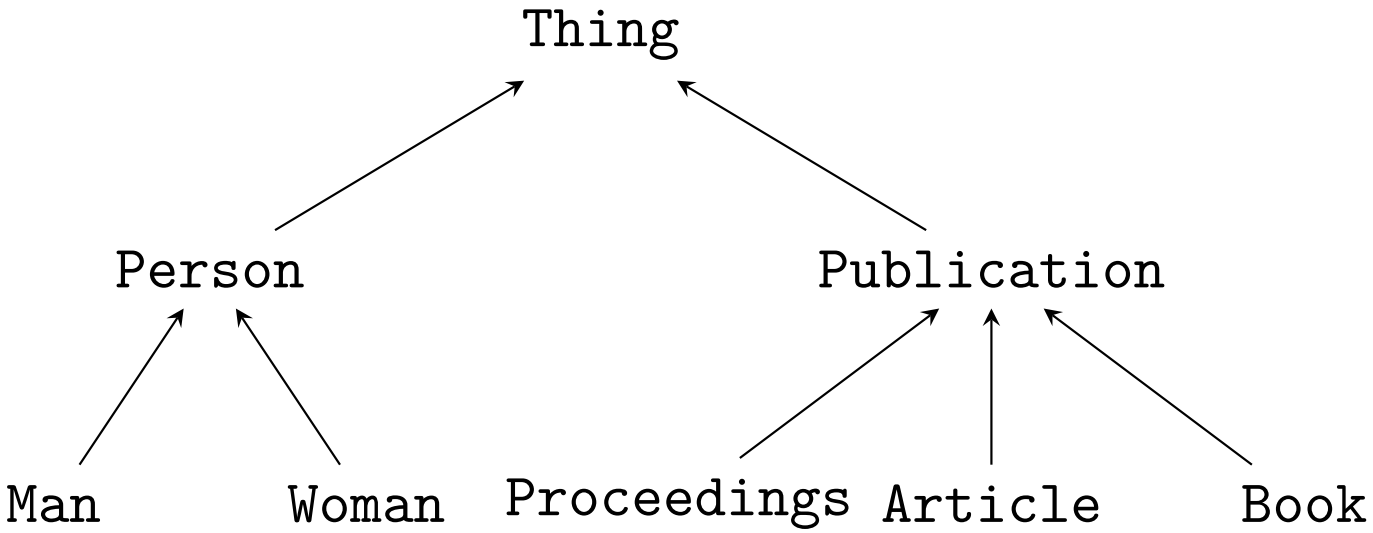
- Well founded possible world semantics.
- Complete expressive power, reasonable concision.
- Possibility of applying queries and updates directly on probabilistic trees, in an efficient way.
- Implementation available.

Complexity study

- Tree-pattern queries are **PTIME**.
- **PTIME** insertions, but exponential-time deletions.

Conceptual model

- IsA ontology of concepts (simple acyclic graph)



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

Semantic representation of a service

A service is described by:

- A tuple of typed input parameters.
- A nested type for its output.
- Semantic relations between inputs and outputs (Datalog-like description).

Services and queries

Exemple

Service giving authors from a publication title:

$$\langle A \rangle \leftarrow \texttt{AuthorOf}(A,P), \texttt{HasTitle}(P,T), \mathbf{Input}(T)$$

Exemple

Query:

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

HTML form analysis

Analyze the structure of HTML forms.

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

Problem

Map relevant domain concepts to each form field.

First step: structural analysis

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

Second step: confirmation with probing

For each field annotated with concept c:

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

Practically speaking, very good precision and good recall; but some limitations on the type of forms that can be dealt with.

Second step: confirmation with probing

For each field annotated with concept c:

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

Practically speaking, very good precision and good recall; but some limitations on the type of forms that can be dealt with.

Query result pages

Extract data from Web pages resulting from a query.

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

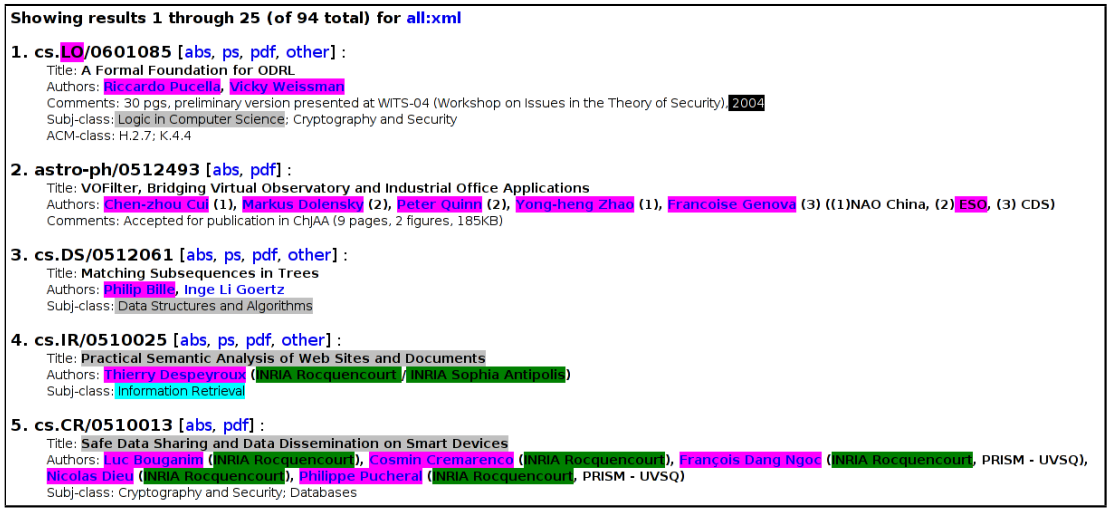
5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt,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

Problems

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

Unsupervised wrapper induction

- Automatic pre-annotation of pages with domain knowledge (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 in a structural information extractor (conditional random fields techniques)

Experimental results

- 10 publication database services.
- Domain knowledge extracted from DBLP.

	Title		Author		Date	
	F_g	F_x	F_g	F_x	F_g	F_x
Average	44	63	64	70	85	76

- F_g: F-measure (in %) of the annotation with domain knowledge.
- F_x: F-measure (in %) of the annotation by the learned wrapper.

Motivation

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

Abstraction

Given two database instances I and J on different schemata, find the optimal description Σ of J given I (with Σ a finite set of formulas).

What does *optimal* mean?

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

Exemple (Tuple-generating dependancies)

R
a
b
c
d

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

$$\Sigma_0 = \varnothing$$

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

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

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

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

Exemple (Tuple-generating dependancies)

R
a
b
c
d

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

$$\Sigma_0 = \varnothing$$

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

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

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

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

Exemple (Relevance computation)

R
a
b
c
d

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

$$\forall x \; R(x) \rightarrow R'(x, x)$$

$R'_{\text{predicted}}$	
a	a
b	b
c	c
d	d

Exemple (Relevance computation)

R
a
b
c
d

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

$$\forall x\ R(x) \wedge x \neq c \rightarrow R'(x, x)$$

$R'_{\text{predicted}}$	
a	a
b	b
d	d

Exemple (Relevance computation)

R
a
b
c
d

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

$$\forall x\ R(x) \wedge x \neq c \rightarrow R'(x, x)$$

$$R'(c, a)$$

$R'_{\text{predicted}}$	
a	a
b	b
c	a
d	d

Exemple (Relevance computation)

R
a
b
c
d

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

$$\forall x\ R(x) \wedge x \neq c \rightarrow R'(x, x)$$

$$R'(c, a)$$

$$R'(g, h)$$

$R'_{\text{predicted}}$	
a	a
b	b
c	c
d	d
g	h

Exemple (Relevance computation)

R
a
b
c
d

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

$$\forall x\ R(x) \wedge x \neq c \rightarrow R'(x,x)$$
$$\exists x \exists y\ R'(x,y) \wedge x = c \wedge y = a$$
$$\exists x \exists y\ R'(x,y) \wedge x = g \wedge y = h$$

$R'_{\text{predicted}}$	
a	a
b	b
c	c
d	d
g	h

Exemple (Relevance computation)

R
a
b
c
d

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

$$\forall x\ R(x) \wedge x \neq c \rightarrow R'(x, x)$$
$$\exists x \exists y\ R'(x, y) \wedge x = c \wedge y = a$$
$$\exists x \exists y\ R'(x, y) \wedge x = g \wedge y = h$$

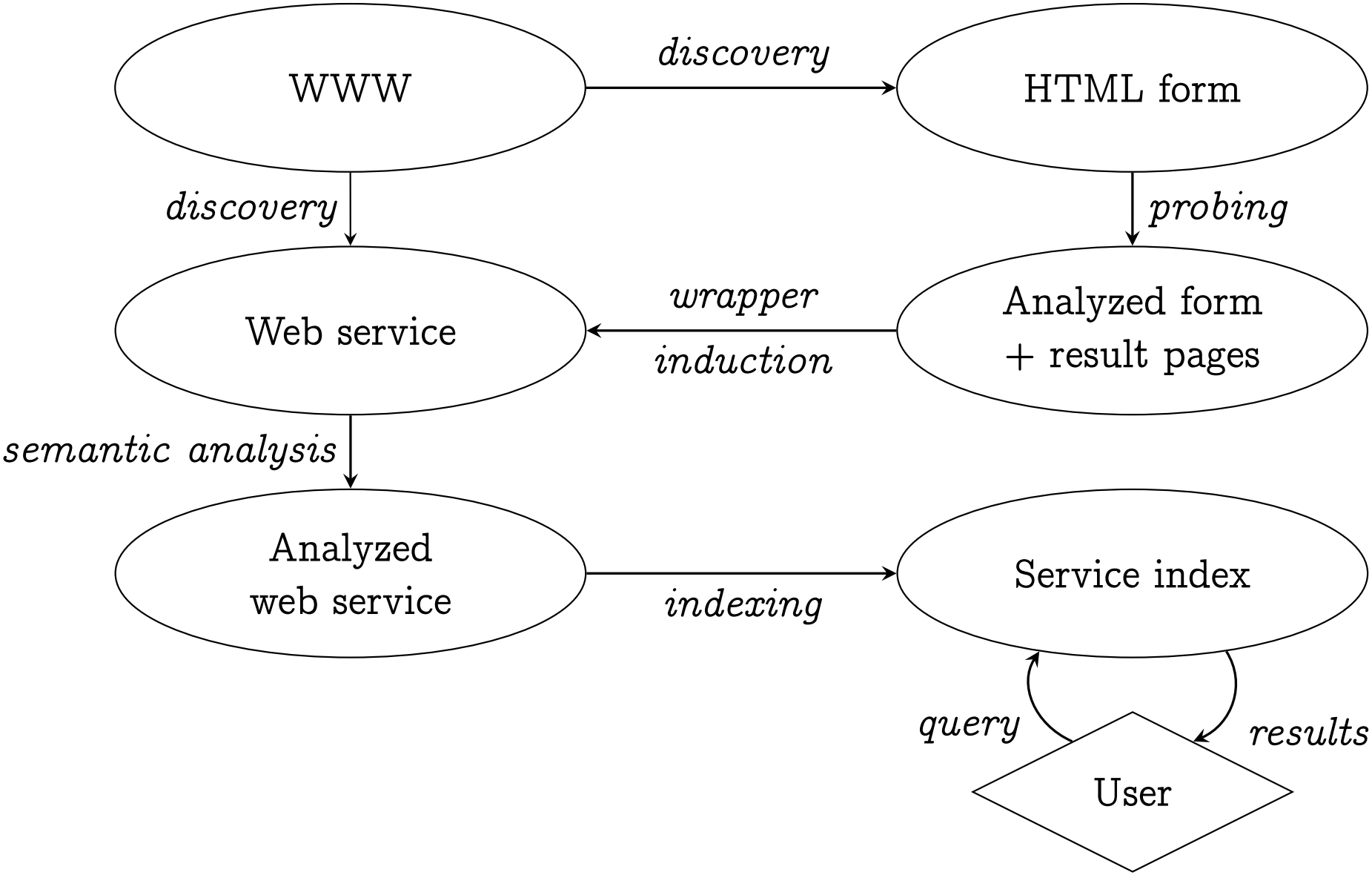
Relevance: 17

$R'_{\text{predicted}}$	
a	a
b	b
c	c
d	d
g	h

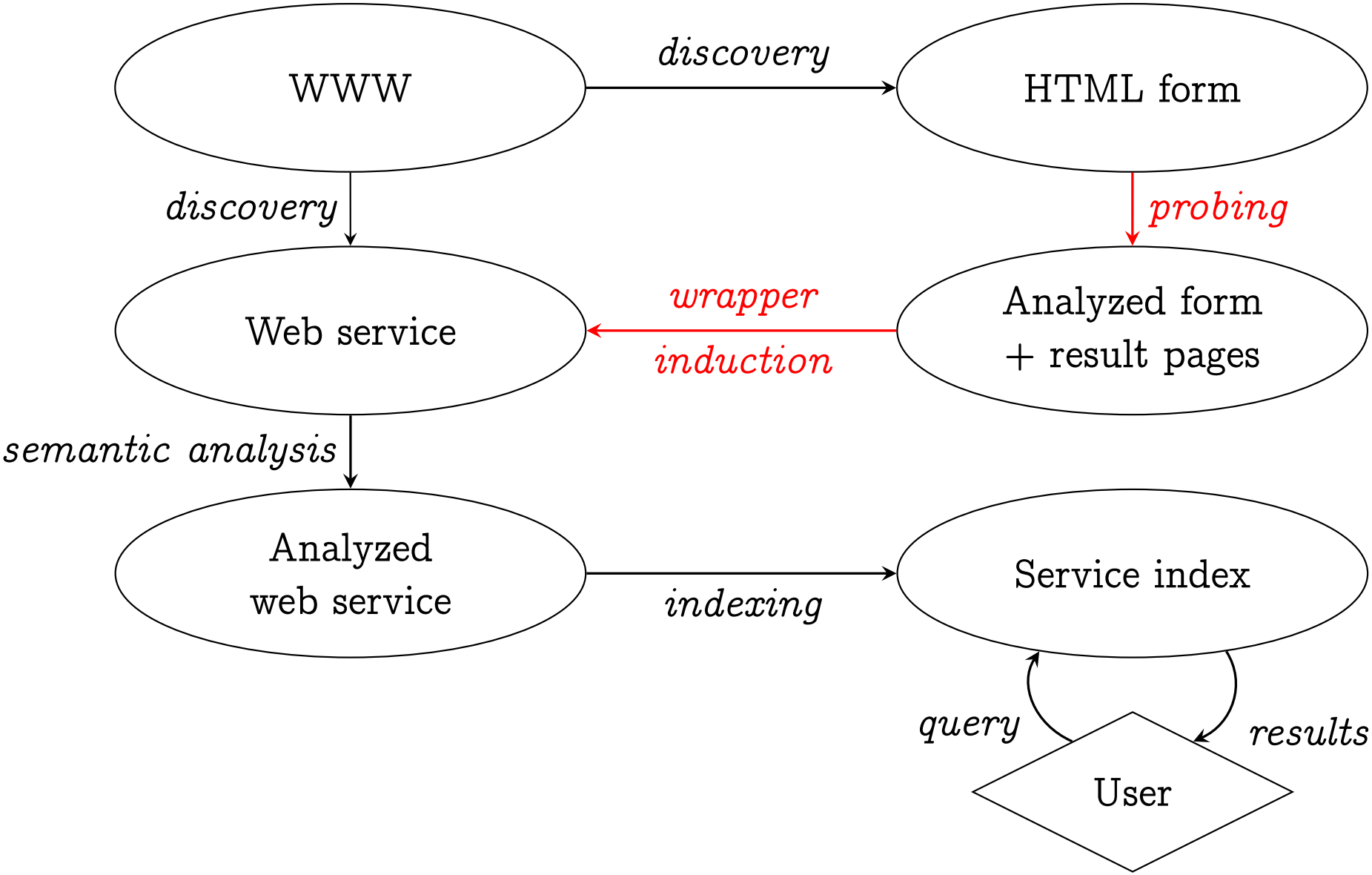
Results

- Description based on the minimal size of a repair of a formula that is valid and explain all facts of J.
- This optimality notion gives "intuitive" results for instances derived from each other with elementary operations.
- Detailed analysis of the algorithmic complexity for various logical languages. High in the polynomial hierarchy (up to $\Pi_4^{\boldsymbol{P}}$ for the optimality of a tuple-generating dependency!).
- Even for $\forall x_1 \forall x_2 \forall x_3\ R(x_1, x_2, x_3) \rightarrow R'(x_1)$, relevance computation is already **NP**-complete.

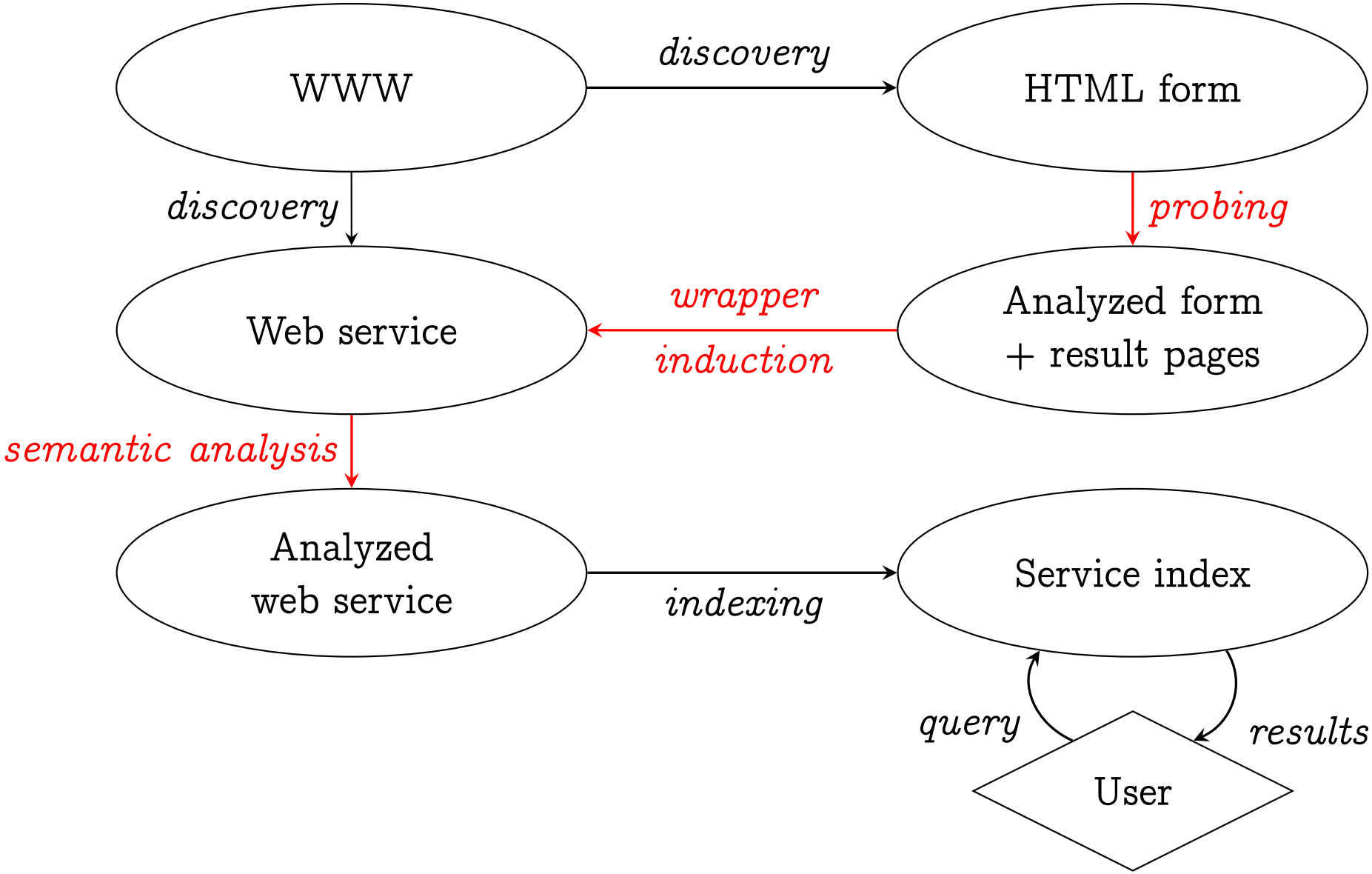
Process for the semantic interpretation of the hidden Web



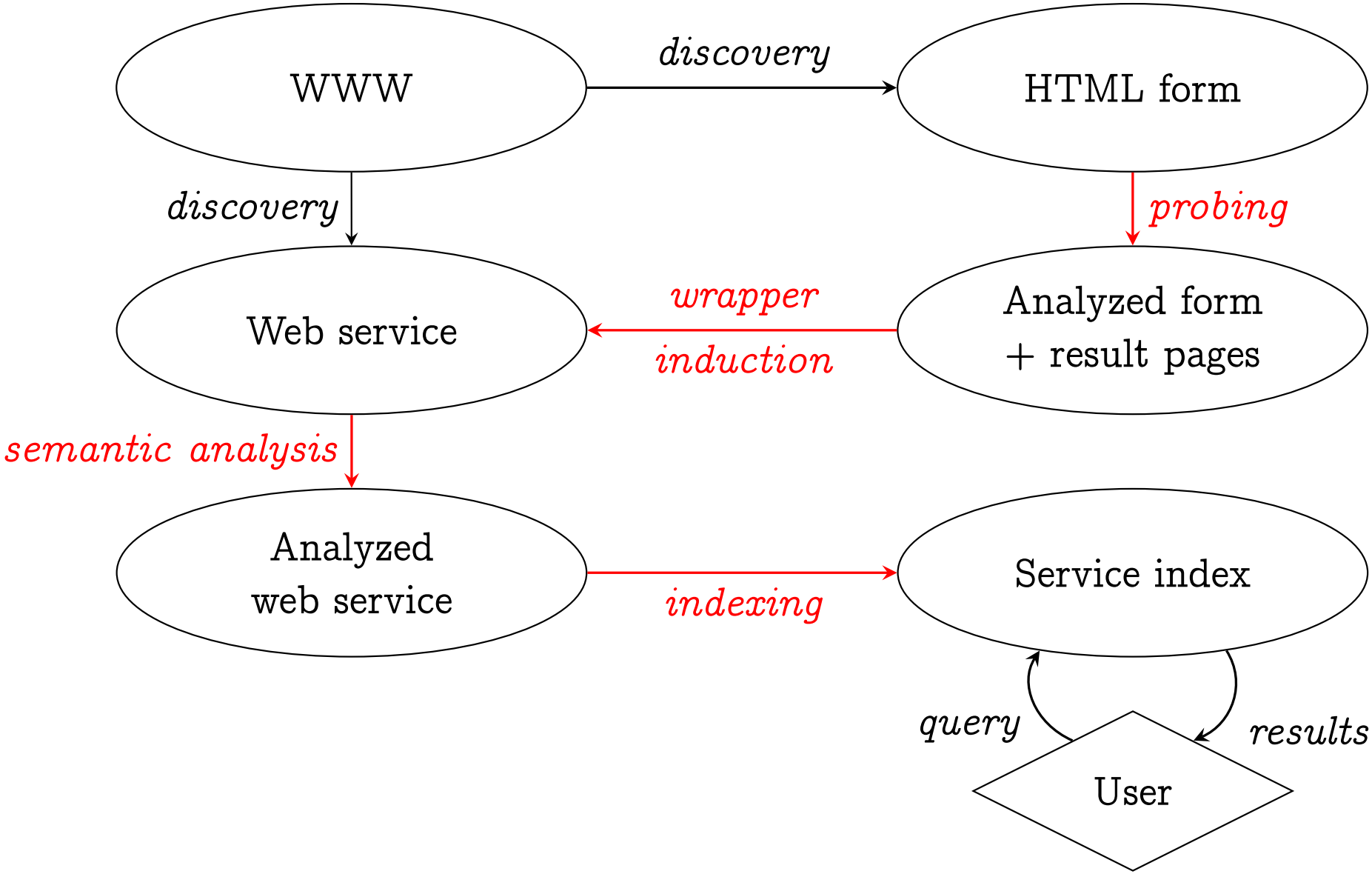
Process for the semantic interpretation of the hidden Web



Process for the semantic interpretation of the hidden Web



Process for the semantic interpretation of the hidden Web



In Summary

1. A general framework for the automatic understanding of the hidden Web, with especially an unsupervised way to discover the structure of a form and of result pages.
2. A semi-structured probabilistic model allowing for queries and updates.
3. A theoretical framework and complexity study for the discovery of schema mappings, solely based on constants appearing in database instances.

Work in progress

- Integration of all the components to obtain a complete system for processing the hidden Web.
- Relation between schema mapping discovery and inductive logic programming.
- Answering queries using views on the semantic model.

Indexing and querying

Given a query, represented as a semantic Web service, how to know which services to query?

Problems

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

Differences with classical databases

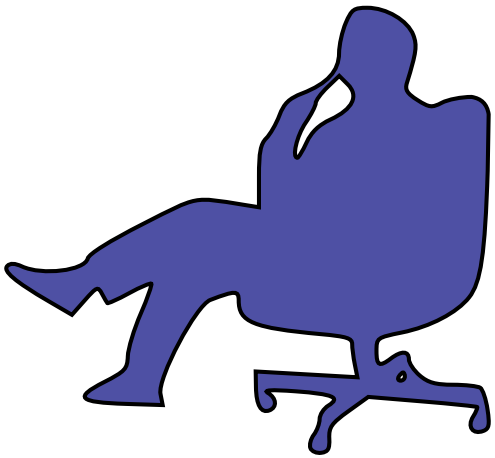
Three main differences:

- Information can only be obtained by views (Local As View).
- Information incomplete and imprecise.
- Nested types.

Three sources of complexity!

Current research directions: Magic sets, Bucket, MiniCon...

Perspectives



- Top-k probabilistic results.
- A machine learning framework adapted to an imperfect annotation, that avoids overfitting (minimal-length description?).
- Deduplication, identification of coreferences despite slightly different information.
- Corroboration of informations between sources.
- Elaborated semantic analysis, based on an understanding of natural language text.

Merci.