

Comprendre le Web caché

Pierre Senellart

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



UNIVERSITÉ
PARIS-SUD 11

Soutenance de thèse de doctorat, 12 décembre 2007



Exemple hypothétique

À l'occasion d'une remise de prix, les amis de Serge décident d'organiser une fête en invitant tous les gens ayant travaillé avec lui.

C'est simple! Il leur suffit de :

- Trouver tous ses coauteurs.
- Pour chacun d'entre eux, trouver leur email actuel.



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	<input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/>	10 results ▾	<input type="button" value="Search Scholar"/>
Author	Return articles written by	<input type="text"/> e.g., "PJ Hayes" or McCarthy		
Publication	Return articles published in	<input type="text"/> e.g., J Biol Chem or Nature		
Date	Return articles published between	<input type="text"/> — <input type="text"/> e.g., 1996		
Subject Areas	<input checked="" type="radio"/> Return articles in all subject areas. <input type="radio"/> Return only articles in the following subject areas: <ul style="list-style-type: none"> <input type="checkbox"/> Biology, Life Sciences, and Environmental Science <input type="checkbox"/> Business, Administration, Finance, and Economics <input type="checkbox"/> Chemistry and Materials Science <input type="checkbox"/> Engineering, Computer Science, and Mathematics <input type="checkbox"/> Medicine, Pharmacology, and Veterinary Science <input type="checkbox"/> Physics, Astronomy, and Planetary Science <input type="checkbox"/> Social Sciences, Arts, and Humanities 			



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	<input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text" value="anywhere in the article"/>	10 results ▾	<input type="button" value="Search Scholar"/>
Author	Return articles written by	<input type="text" value="abiteboul"/> e.g., "PJ Hayes" or McCarthy		
Publication	Return articles published in	<input type="text"/> e.g., J Biol Chem or Nature		
Date	Return articles published between	<input type="text"/> — <input type="text"/> e.g., 1996		
Subject Areas	<input checked="" type="radio"/> Return articles in all subject areas. <input type="radio"/> Return only articles in the following subject areas:			
	<input type="checkbox"/> Biology, Life Sciences, and Environmental Science <input type="checkbox"/> Business, Administration, Finance, and Economics <input type="checkbox"/> Chemistry and Materials Science <input type="checkbox"/> Engineering, Computer Science, and Mathematics <input type="checkbox"/> Medicine, Pharmacology, and Veterinary Science <input type="checkbox"/> Physics, Astronomy, and Planetary Science <input type="checkbox"/> Social Sciences, Arts, and Humanities			



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](#)
[Quervying 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 manipulating 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](#)



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](#)
[Quervying 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 manipulating 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 *

 <p>Inria</p>	<input type="text"/> Nom	Saisir un nom de Projet <input type="text"/> ou sélectionner dans cette liste.
	<input type="text"/> Prénom	
<input type="button" value="Rechercher"/> <input type="button" value="Réinitialisation"/>		

[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](#).



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](#)
[Quervying 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 manipulating 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](#)



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](#)
[Quervying 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 manipulating 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 *

 <p>Inria</p>	<input type="text"/> Nom	Saisir un nom de Projet <input type="text"/> ou sélectionner dans cette liste.
	<input type="text"/> Prénom	
<input type="button" value="Rechercher"/> <input type="button" value="Réinitialisation"/>		


[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 *

 Inria	<input type="text" value="christophides"/> Nom	Saisir un nom de Projet <input type="text"/> ou sélectionner dans cette liste.
	<input type="text"/> Prénom	
<input type="button" value="Rechercher"/> <input type="button" value="Réinitialisation"/>		<ul style="list-style-type: none"> ABS ABSTRACT ABSTRACTION ACACIA ACES ACTIVEEON ADAGE

[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 Courriel électronique Téléphone Site Projet

[Interroger à nouveau l'annuaire Inria](#) [Poster vos questions et commentaires.](#)

Constatations

Fastidieux !

- Identifier tous les **services** pertinents.
- Comprendre la façon de les **interroger**.
- Les interroger **un à un**.
- **Fusionner** les résultats.

Objectif

Faire faire tout ce travail à l'ordinateur, de manière entièrement **automatique**.

Le Web caché

Définition (Web caché, Web profond, Web invisible)

L'ensemble des contenus du Web qui ne sont pas accessibles depuis la **structure d'hyperliens** du World Wide Web. En général : formulaires HTML, services Web.

Estimation de taille (2001) : 500 fois plus de données que le **Web de surface**.

Comment comprendre et bénéficier de ce contenu ?

Comprendre le Web caché

Objectif

- Indexation **en compréhension** du Web caché.
- Requêtes de **haut niveau**.
- ⇒ un moteur de recherche **sémantique** pour le Web caché.

**De manière entièrement automatique
et non supervisée !**

- Problème **difficile** et **vaste**.
- Utilisation de **connaissance du domaine** (**ontologie**, **instances**).
- Exemple du domaine des publications scientifiques.

Comprendre le Web caché

Objectif

- Indexation **en compréhension** du Web caché.
- Requêtes de **haut niveau**.
- ⇒ un moteur de recherche **sémantique** pour le Web caché.

**De manière entièrement automatique
et non supervisée !**

- Problème **difficile** et **vaste**.
- Utilisation de **connaissance du domaine** (**ontologie**, **instances**).
- Exemple du domaine des publications scientifiques.

1 Introduction

2 Cadre général

- Processus général
- Imprécision
- Modèle de données XML probabiliste
- Modèle sémantique

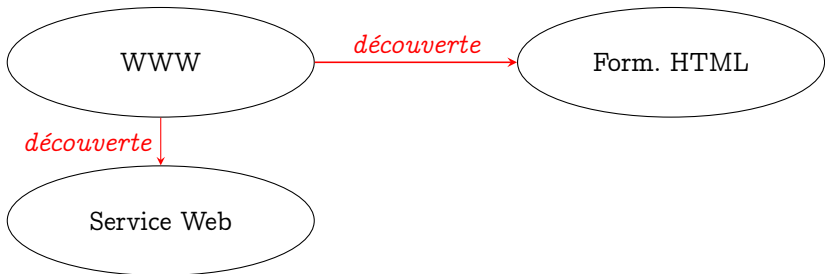
3 Différents modules

4 Conclusion

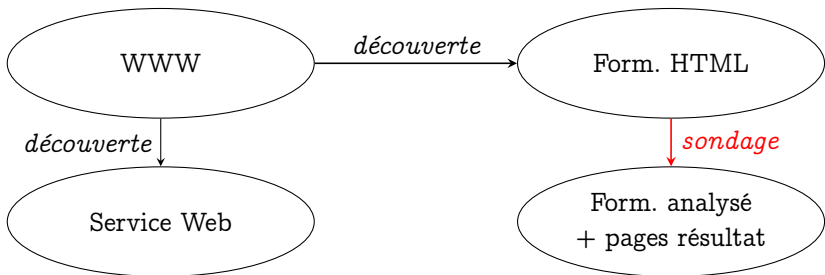
Processus d'interprétation sémantique du Web caché



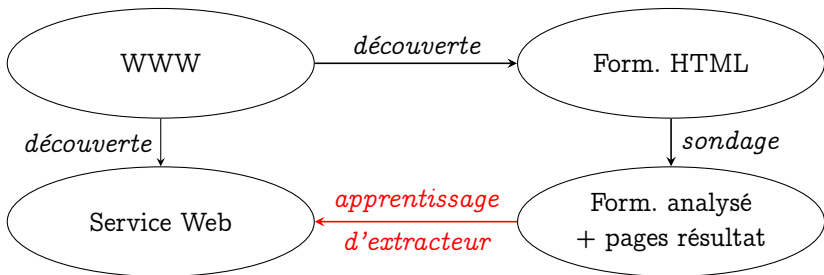
Processus d'interprétation sémantique du Web caché



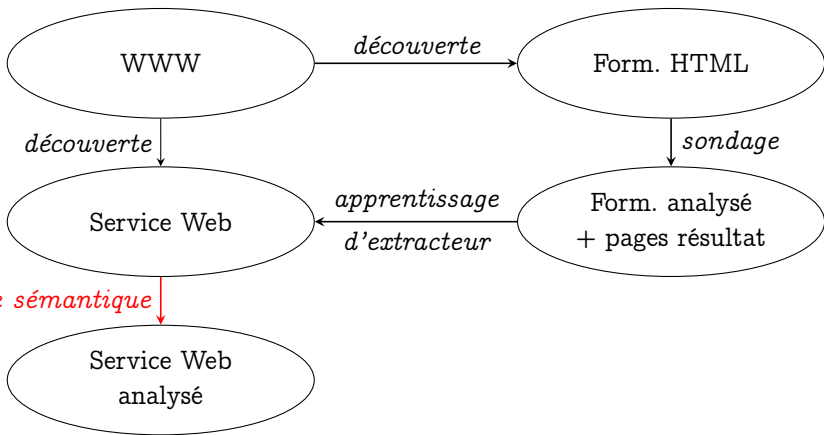
Processus d'interprétation sémantique du Web caché



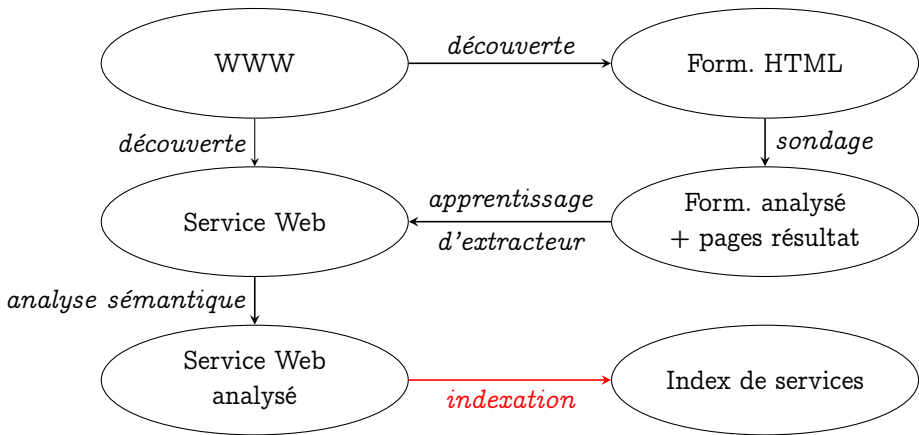
Processus d'interprétation sémantique du Web caché



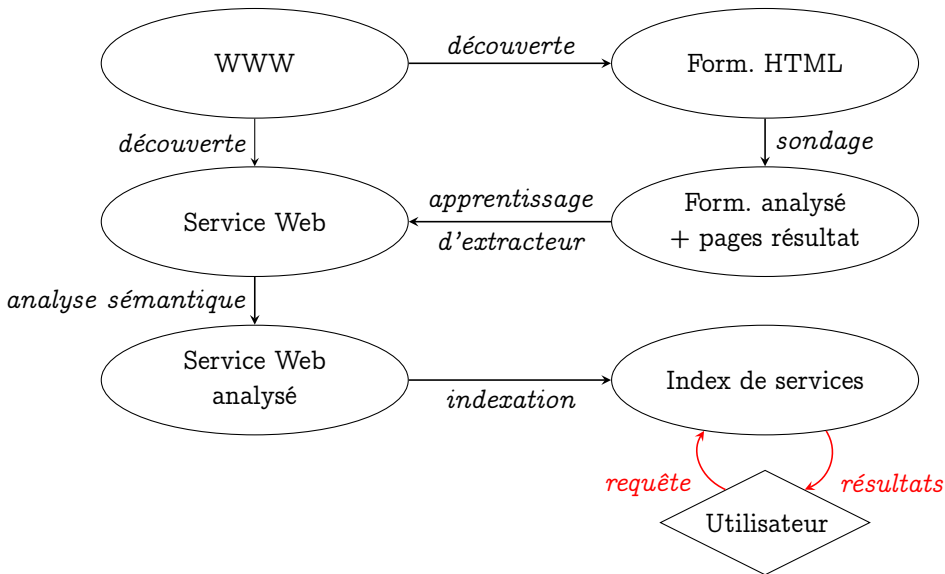
Processus d'interprétation sémantique du Web caché



Processus d'interprétation sémantique du Web caché



Processus d'interprétation sémantique du Web caché

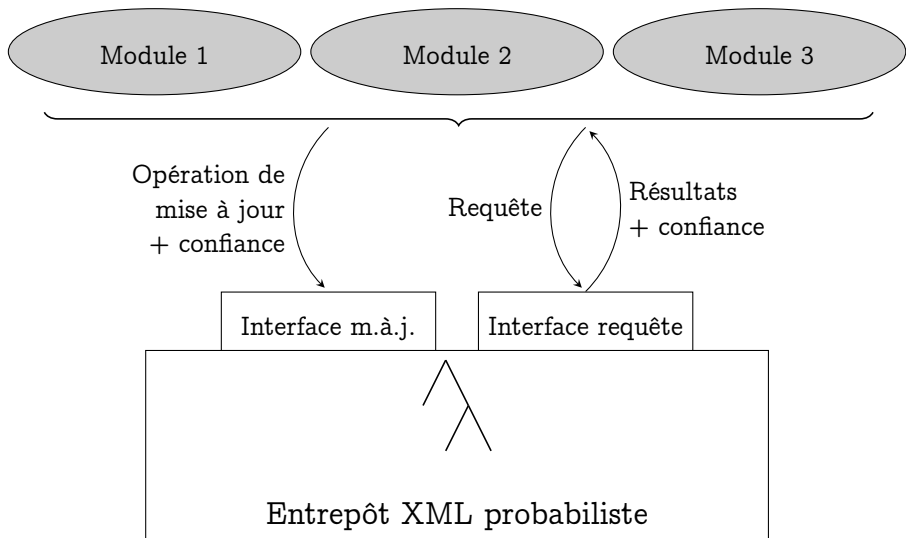


Données et tâches imprécises

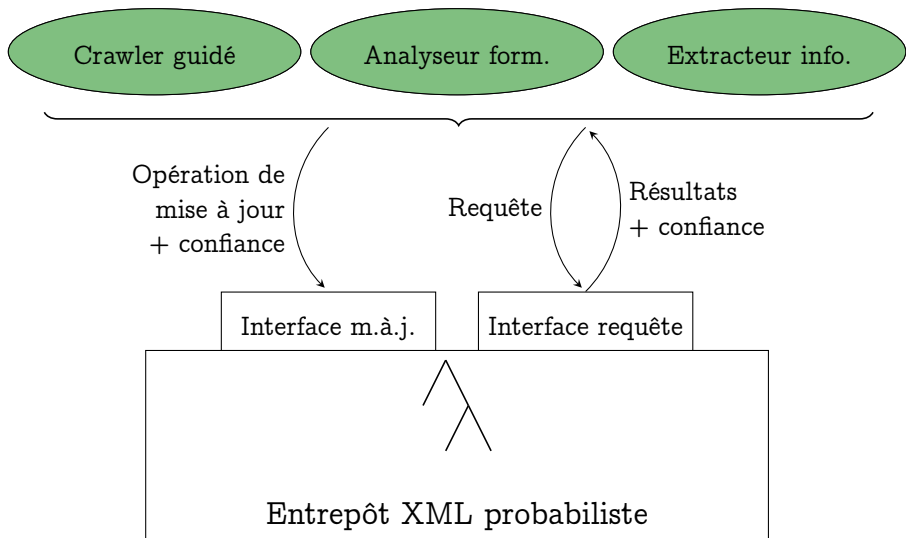
Observations

- Beaucoup des tâches nécessaires génèrent des données **imprécises**, avec une certaine valeur de **confiance**.
- Besoin de gérer cette imprécision, de travailler avec **tout au long d'un processus complexe**.

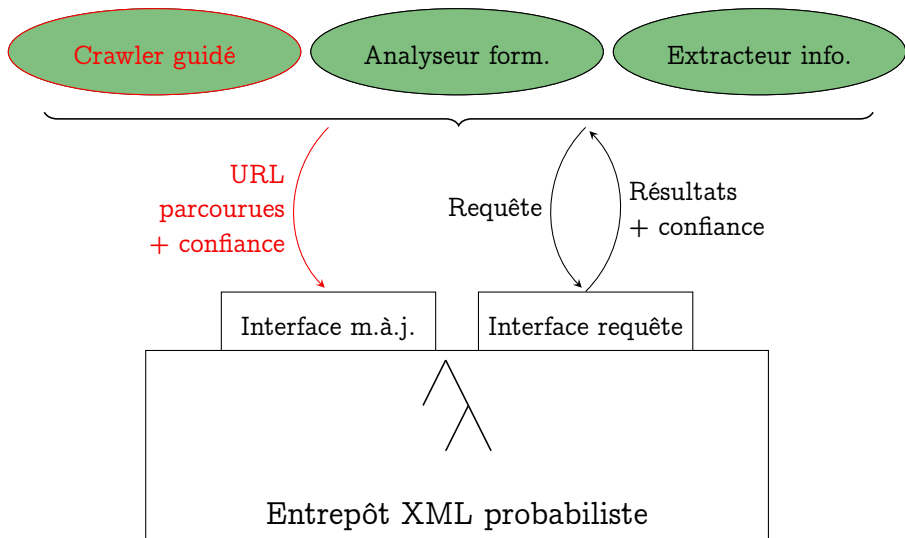
Un entrepôt probabiliste XML



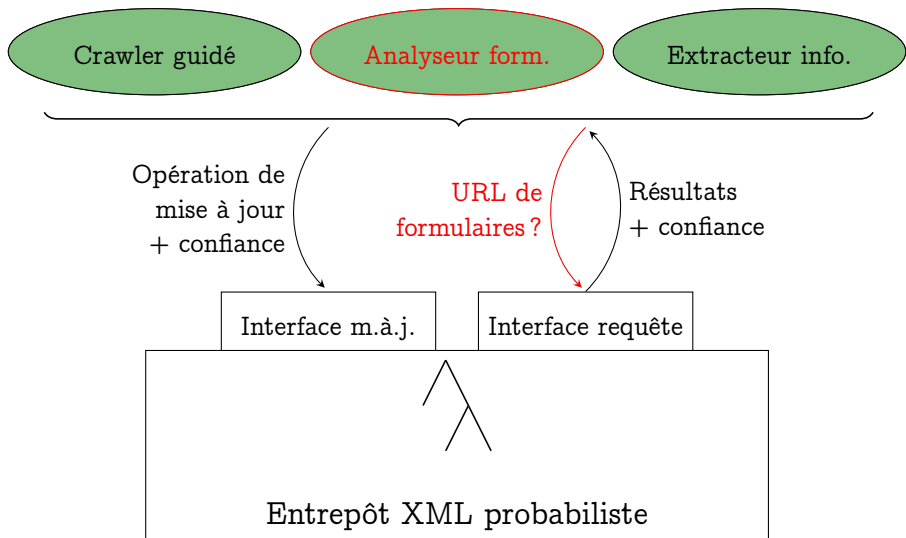
Un entrepôt probabiliste XML (Web caché)



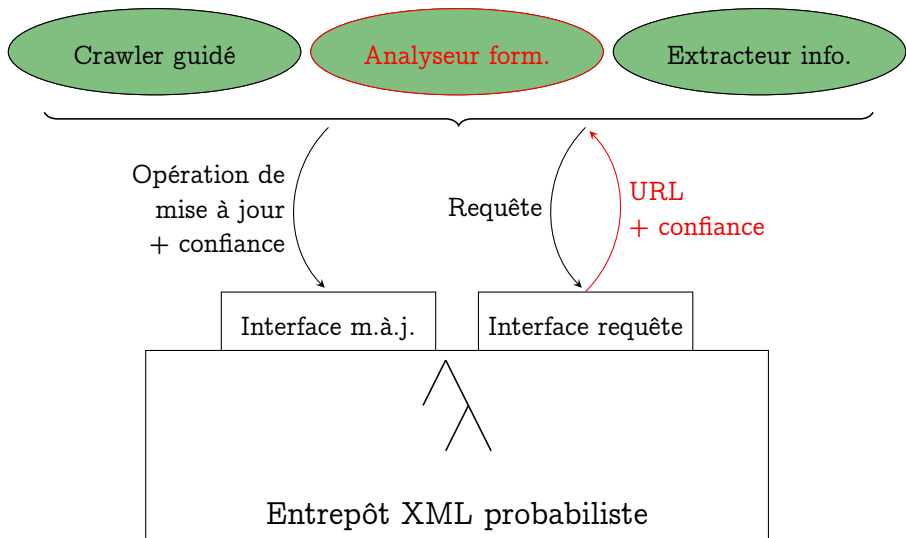
Un entrepôt probabiliste XML (Web caché)



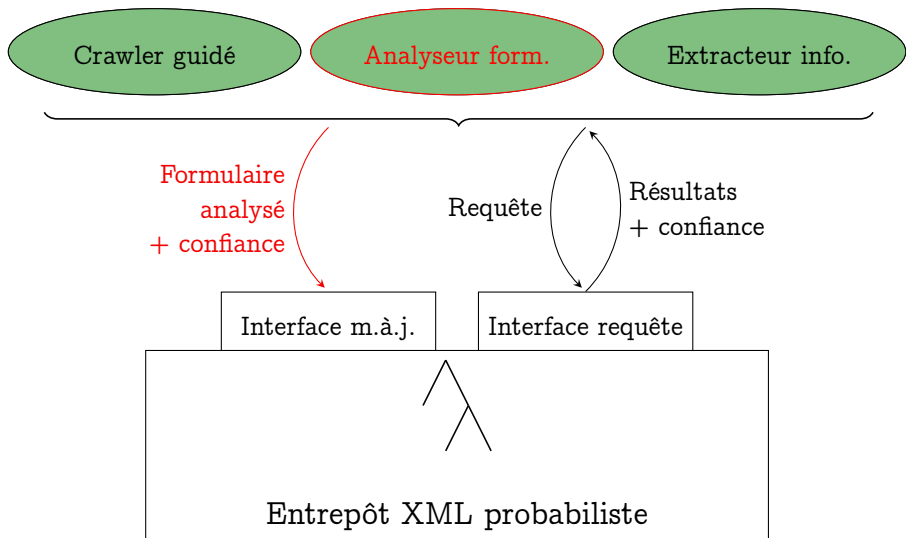
Un entrepôt probabiliste XML (Web caché)



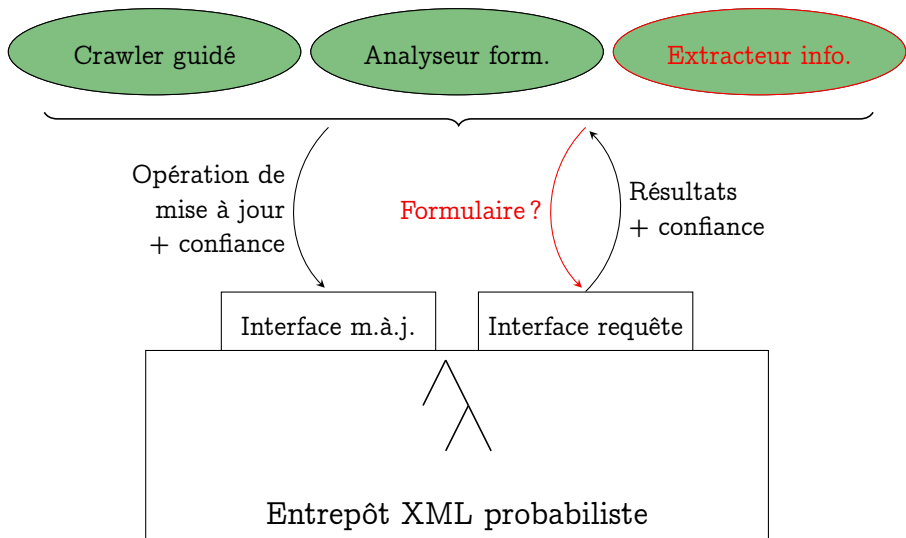
Un entrepôt probabiliste XML (Web caché)



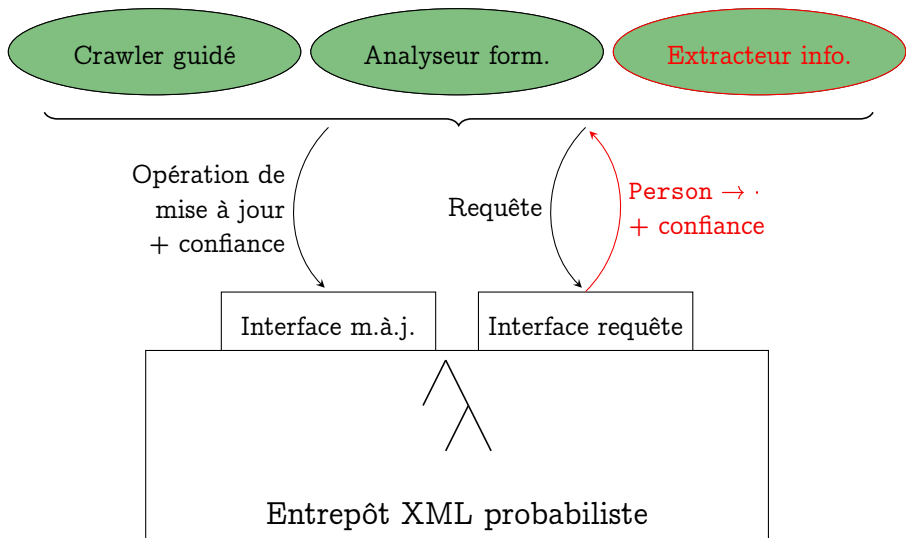
Un entrepôt probabiliste XML (Web caché)



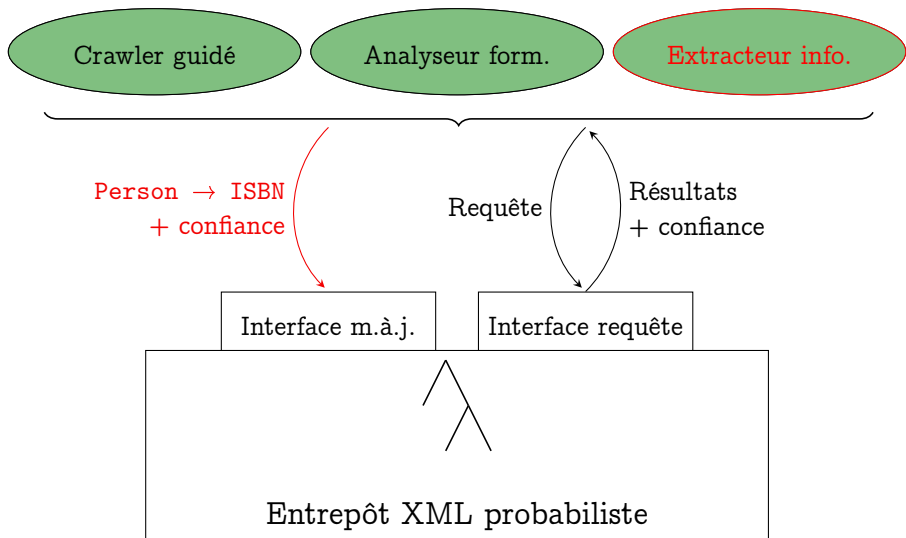
Un entrepôt probabiliste XML (Web caché)



Un entrepôt probabiliste XML (Web caché)

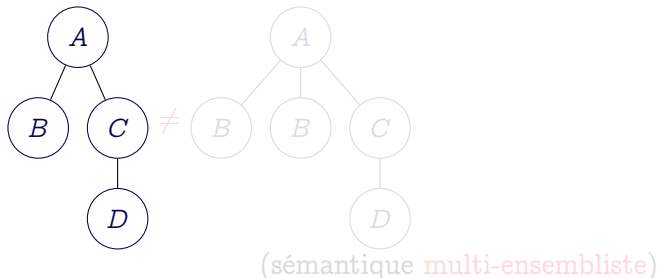


Un entrepôt probabiliste XML (Web caché)



Arbres probabilistes

- Cadre
- Arbres de données **non ordonnés**.
 - Simplifications (sans perte de généralité) : pas d'attributs, pas de contenu mixte...

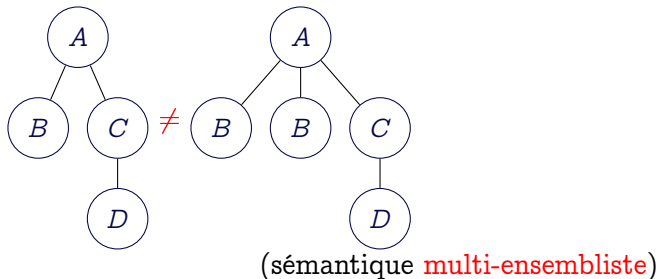


Univers : Ensemble des tels arbres.

Arbre probabiliste : Représentation d'une **distribution discrète de probabilité** dans cet univers.

Arbres probabilistes

- Cadre
- Arbres de données **non ordonnés**.
 - Simplifications (sans perte de généralité) : pas d'attributs, pas de contenu mixte...

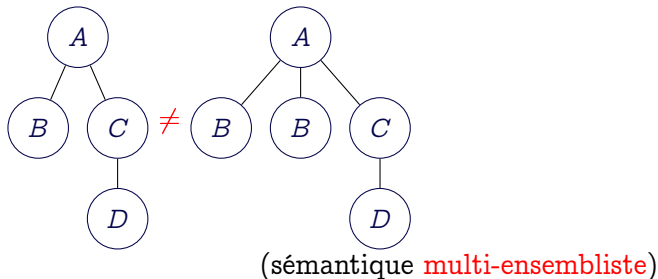


Univers : Ensemble des tels arbres.

Arbre probabiliste : Représentation d'une **distribution discrète de probabilité** dans cet univers.

Arbres probabilistes

- Cadre
- Arbres de données **non ordonnés**.
 - Simplifications (sans perte de généralité) : pas d'attributs, pas de contenu mixte...

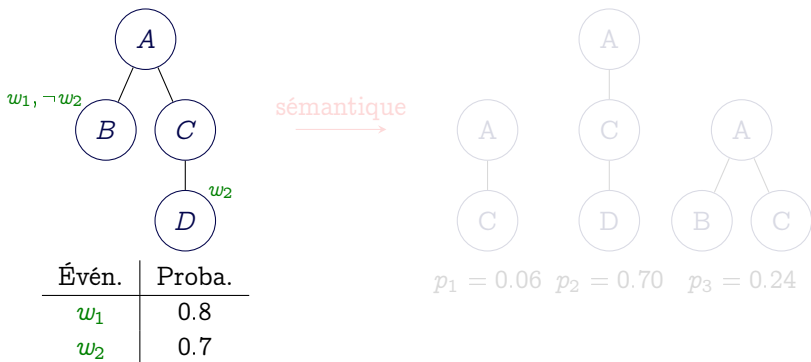


Univers : Ensemble des tels arbres.

Arbre probabiliste : Représentation d'une **distribution discrète de probabilité** dans cet univers.

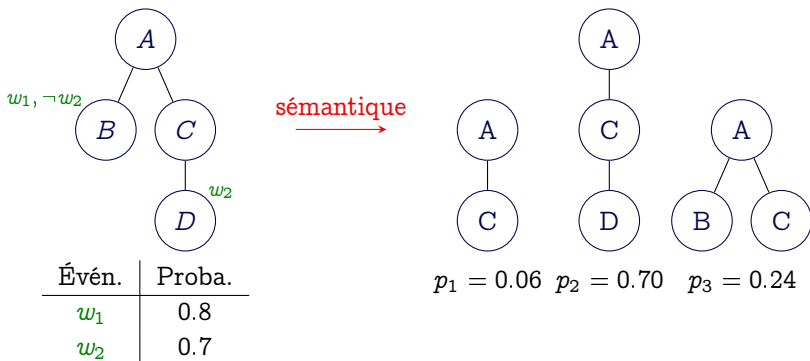
Modèle d'arbre probabiliste

- Arbres de données avec des **conditions d'événements** (conjonction d'événements probabilistes ou de leurs négations) **affectées à chaque nœud**.
- Événements probabilistes : **variables aléatoires booléennes**, supposées indépendantes, avec leur propre distribution de probabilité.



Modèle d'arbre probabiliste

- Arbres de données avec des **conditions d'événements** (conjonction d'événements probabilistes ou de leurs négations) **affectées à chaque nœud**.
- Événements probabilistes : **variables aléatoires booléennes**, supposées indépendantes, avec leur propre distribution de probabilité.



Caractéristiques du modèle d'arbre probabiliste

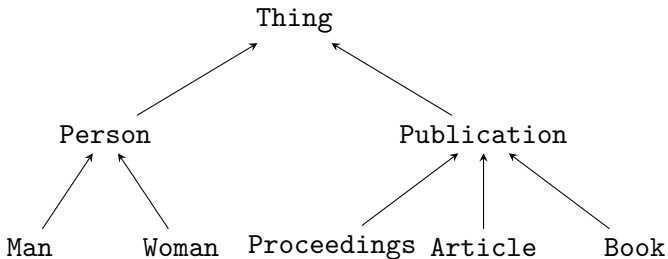
- Sémantique de **mondes possibles** bien définie.
- Pouvoir expressif **complet**, **concision** raisonnable.
- Possibilité d'appliquer **requêtes** et **misés à jour** directement sur les arbres probabilistes, de manière **efficace**.
- **Implémentation** disponible.

Étude de complexité

- Requêtes de motifs d'arbre **P_{TIME}**.
- Insertions **P_{TIME}**, mais suppressions exponentielles.

Modèle conceptuel

- **Ontologie** SorteDe de **concepts** (simple graphe acyclique)



- **Rôles** *n*-aires typés
 - AuthorOf(Publication, Person)
 - HasName(Person, Name)

Représentation sémantique d'un service

Par quoi un service est-il décrit ?

- Un n -uplet de paramètres d'entrées **typés**.
- Un type **imbriqué** pour sa sortie.
- Des relations sémantiques entre entrées et sorties (description à la **Datalog**).

Services et requêtes

Exemple

Service donnant des auteurs à partir d'un titre de publication :

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

Exemple

Requête :

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

- 1 Introduction
- 2 Cadre général
- 3 Différents modules**
 - Analyse des formulaires
 - Analyse des pages de résultats
 - Analyse sémantique des services
- 4 Conclusion

- 1 Introduction
- 2 Cadre général
- 3 Différents modules**
 - Analyse des formulaires
 - Analyse des pages de résultats
 - Analyse sémantique des services
- 4 Conclusion

Analyse des formulaires HTML

Analyser la **structure** de formulaires HTML.

Authors	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Title	<input type="text"/>		Year <input type="text"/>	Page <input type="text"/>
Conference	<input type="text"/>		ID	<input type="text"/>
Journal	<input type="text"/>		Volume <input type="text"/>	Number <input type="text"/>
<input type="button" value="Search"/>	<input type="button" value="Reset"/>	Maximum of <input type="text" value="100"/> matches		

Problème

Associer à chaque champ de formulaire pertinent le **concept du domaine** approprié.

Première étape : analyse structurelle

- 1 Construire un **contexte** pour chaque champ :
 - élément label ;
 - attributs id et name ;
 - texte précédent immédiatement le champ.
- 2 Enlever les **mots grammaticaux**, **lemmatiser**.
- 3 **Apparier** ce contexte avec les noms de concepts, étendus avec WordNet.
- 4 Obtenir ainsi des **annotations candidates**.

Deuxième étape : confirmation par sondage

Pour chaque champ annoté avec un concept c :

- 1 Sonder le champ avec un mot absurde pour produire une **page d'erreur**.
- 2 **Sonder** le champ avec des instances de c (choisies représentatives de la distribution de fréquence de c).
- 3 Comparer les pages obtenues avec la page d'erreur (en utilisant une classification suivant la structure d'arbre), pour distinguer les pages d'erreurs des **pages de résultat**.
- 4 **Confirmer** l'annotation si suffisamment de pages de résultat ont été obtenues.

En pratique, **très bonne précision** et **bon rappel** ; mais des limitations sur le type de formulaires qui peut être traité.

Deuxième étape : confirmation par sondage

Pour chaque champ annoté avec un concept c :

- 1 Sonder le champ avec un mot absurde pour produire une **page d'erreur**.
- 2 **Sonder** le champ avec des instances de c (choisies représentatives de la distribution de fréquence de c).
- 3 Comparer les pages obtenues avec la page d'erreur (en utilisant une classification suivant la structure d'arbre), pour distinguer les pages d'erreurs des **pages de résultat**.
- 4 **Confirmer** l'annotation si suffisamment de pages de résultat ont été obtenues.

En pratique, **très bonne précision** et **bon rappel** ; mais des limitations sur le type de formulaires qui peut être traité.

- 1 Introduction
- 2 Cadre général
- 3 Différents modules**
 - Analyse des formulaires
 - Analyse des pages de résultats
 - Analyse sémantique des services
- 4 Conclusion

Pages de résultat à une requête

Extraire des données des pages Web de résultat à une requête.

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

- 1. cs.LO/0601085 [abs, ps, pdf, other] :**
Title: A Formal Foundation for OORL
Authors: [Ricardo 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: VOFitter, 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 Cremerence](#) (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

Problèmes

- **Quelle partie** de la page Web contient la réponse ?
- Comment extraire du **contenu structuré** ?

Apprentissage non supervisé d'extracteur

- Pré-annoter automatiquement les pages avec la connaissance du domaine (techniques d'automates finis) : à la fois **imparfait** et **incomplet**.

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

- 1. cs.LG/0601085 [abs, ps, pdf, other] :**
 Title: **A Formal Foundation for ODRL**
 Authors: **Rafael Peeters**, **Weng-Wee Koh**
 Comments: 30 pgs. preliminary version presented at WTS-04 (Workshop on Issues in the Theory of Security) 2006
 Sub-class: **Logic in Computer Science**: Cryptography and Security
 ACM-class: H.2.7; K.4.4
- 2. astro-ph/0512493 [abs, pdf] :**
 Title: **VDFilter, Bridging Virtual Observatory and Industrial Office Applications**
 Authors: **Yongchang Gao** (1), **Yongqiang Gao** (2), **Baojun Zhang** (1), **Yanbinbin Li** (3) (1)IAAO China, (2)CS, (3) CDS
 Comments: Accepted for publication in CSPA (9 pages, 2 figures, 195KB)
- 3. cs.DS/0512061 [abs, ps, pdf, other] :**
 Title: **Matching Subsequences in Trees**
 Authors: **Hajin Park**, **Inge Li Goertz**
 Sub-class: **Data Structures and Algorithms**
- 4. cs.IR/0510025 [abs, ps, pdf, other] :**
 Title: **Practical Semantic Analysis of Web Sites and Documents**
 Authors: **Frederik Deputat**, **Andreas Kerschbaum**, **Frank S. Roth**
 Sub-class: **Information Retrieval**
- 5. cs.CR/0510013 [abs, pdf] :**
 Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
 Authors: **Eugene Sionis**, **Georgios Christodoulopoulos**, **Francisco Dana Nader**, **Christoforos Demetrescu**, **Nicolas Dasi**, **Christoforos Demetrescu**, **Matthias Fischer**, **Christoforos Demetrescu**, **PRISM - UVSQ**
 Subclass: **Cryptography and Security**: Databases

- Utiliser de l'**apprentissage** pour **généraliser** le résultat en un extracteur structurel d'information (techniques de champs aléatoires conditionnels).

Résultats expérimentaux

- Une dizaine de services de bases de données de publications.
- Connaissance de domaine extraite de DBLP.

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

- F_g : F -measure (en %) de l'annotation par la connaissance du domaine.
- F_x : F -measure (en %) de l'annotation par l'extracteur appris.

- 1 Introduction
- 2 Cadre général
- 3 Différents modules**
 - Analyse des formulaires
 - Analyse des pages de résultats
 - Analyse sémantique des services
- 4 Conclusion

Motivation

Analyser les **relations** entre des sources différentes, ou entre une source et la connaissance du domaine.

Abstraction

Étant données deux instances de bases de données I et J sur des schémas différents, trouver la description **optimale** Σ de J sachant I (avec Σ un ensemble fini de formules).

Que signifie *optimal* ?

- **Concision** de la description.
- **Validité** des faits prédits par I et Σ .
- Faits de J **expliqués** par I et Σ .

Exemple (Dépendances génératrices de n -uplets)

<u>R</u>	<u>R'</u>
a	a a
b	b b
c	c a
d	d d
	g h

$$\Sigma_0 = \emptyset$$

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

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

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

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

Exemple (Dépendances génératrices de n -uplets)

<u>R</u>	<u>R'</u>
a	a a
b	b b
c	c a
d	d d
	g h

$$\Sigma_0 = \emptyset$$

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

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

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

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

Exemple (Calcul de pertinence)

<u>R</u>
a
b
c
d

<u>R'</u>
a a
b b
c a
d d
g h

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

<u>$R'_{\text{prédit}}$</u>
a a
b b
c c
d d

Exemple (Calcul de pertinence)

<u>R</u>
a
b
c
d

<u>R'</u>
a a
b b
c a
d d
g h

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

<u>$R'_{\text{prédit}}$</u>
a a
b b
d d

Exemple (Calcul de pertinence)

<u>R</u>
a
b
c
d

<u>R'</u>
a a
b b
c a
d d
g h

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

<u>$R'_{\text{prédit}}$</u>
a a
b b
c a
d d

Exemple (Calcul de pertinence)

<u>R</u>
a
b
c
d

<u>R'</u>
a a
b b
c a
d d
g h

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

$$R'(c, a)$$

$$R'(g, h)$$

<u>$R'_{\text{prédit}}$</u>
a a
b b
c c
d d
g h

Exemple (Calcul de pertinence)

R
a
b
c
d

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

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

$$\exists x \exists y R'(x, y) \wedge x = c \wedge y = a$$

$$\exists x \exists y R'(x, y) \wedge x = g \wedge y = h$$

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

Exemple (Calcul de pertinence)

R
a
b
c
d

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

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

$$\exists x \exists y R'(x, y) \wedge x = c \wedge y = a$$

$$\exists x \exists y R'(x, y) \wedge x = g \wedge y = h$$

Pertinence : 17

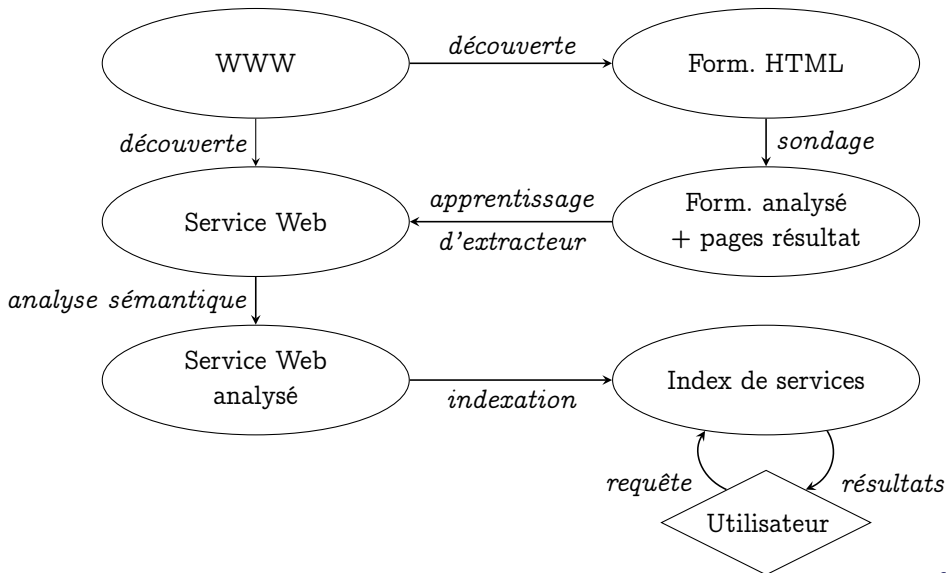
$R'_{\text{prédit}}$
a a
b b
c c
d d
g h

Résultats

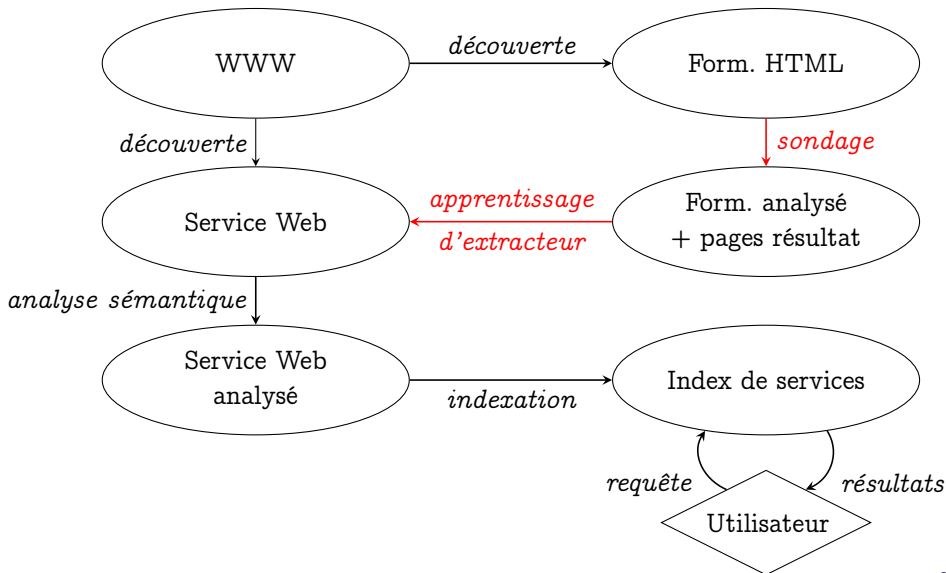
- Description basée sur la **taille minimale** d'une réparation d'une formule valide et expliquant tous les faits de J .
- Cette notion d'optimalité donne des résultats « intuitifs » pour des instances dérivées à partir d'opérations élémentaires.
- Analyse détaillée de **complexité algorithmique** pour différents langages logiques. Élevée dans la hiérarchie polynomiale (jusqu'à Π_4^P pour l'optimalité d'une dépendance génératrice de n -uplets!).
- Même pour $\forall x_1 \forall x_2 \forall x_3 R(x_1, x_2, x_3) \rightarrow R'(x_1)$, le calcul de la pertinence d'une formule est déjà **NP-complet**.

- 1 Introduction
- 2 Cadre général
- 3 Différents modules
- 4 Conclusion**
 - Le Web caché
 - Autres travaux

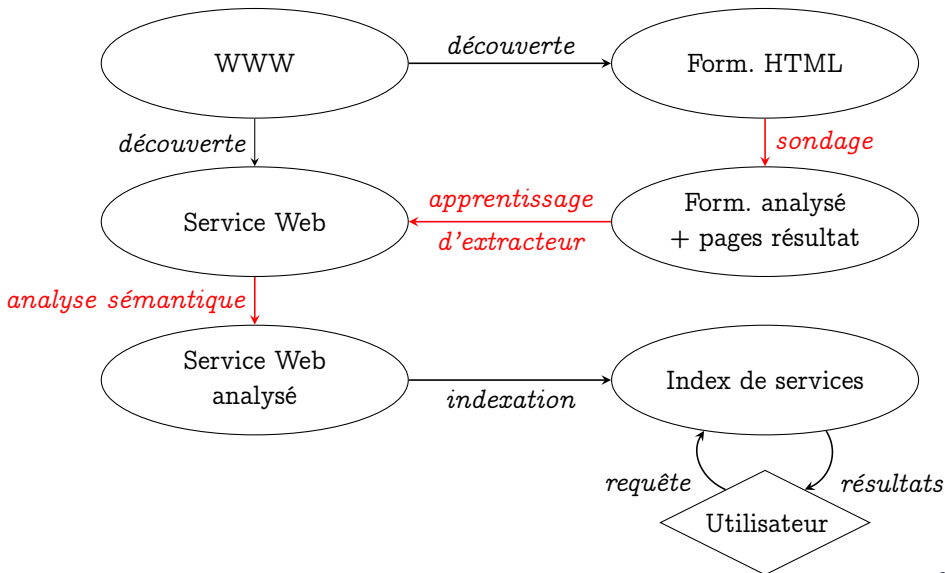
Processus d'interprétation sémantique du Web caché



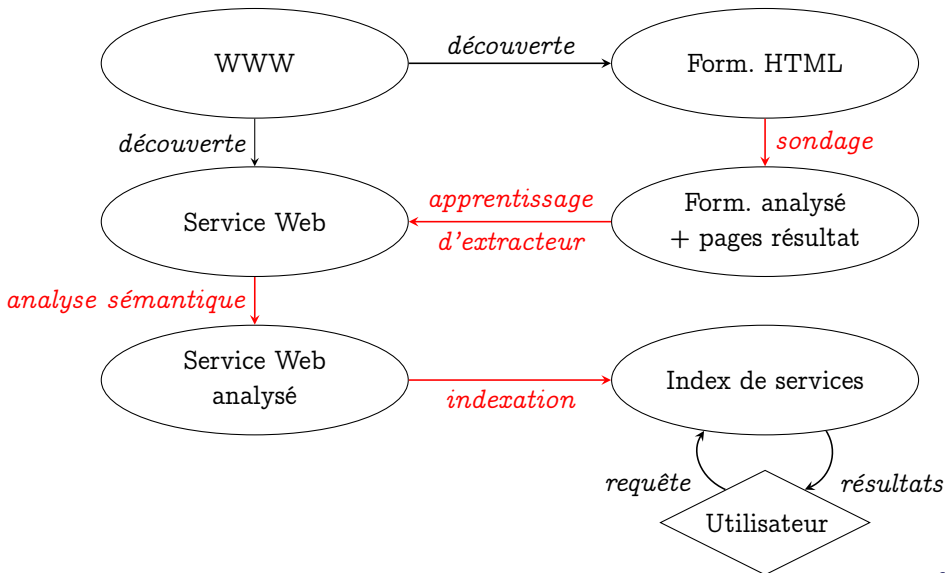
Processus d'interprétation sémantique du Web caché



Processus d'interprétation sémantique du Web caché



Processus d'interprétation sémantique du Web caché



Résumé des contributions

- ① Un cadre général pour la compréhension **automatique** du **Web caché**, avec en particulier une manière non supervisée de découvrir la structure d'un formulaire et de pages de résultats [Soumis WWW 2008a].
- ② Un **modèle probabiliste semi-structuré** permettant requêtes et mises à jour, avec implémentation et étude de complexité [EDBT 2006, PODS 2007].
- ③ Un cadre théorique et une étude de complexité pour la **découverte de correspondances de schémas**, en se basant uniquement sur les constantes apparaissant dans des instances de bases de données [Soumis PODS 2008].

Travaux en cours



- **Intégration** de tous les composants pour obtenir un système complet de traitement du Web caché.
- Relation entre découverte de correspondances de schémas et **programmation logique inductive**.
- **Réponse à des requêtes via des vues** sur le modèle sémantique.

Indexation et interrogation

Étant donné une **requête**, représentée comme un service Web sémantique, comment savoir quels services interroger ?

Problèmes

- **Subsomption** de paramètres d'entrée/sortie.
- Paramètres d'entrée **manquants**.
- **Composition** de services Web.

Différences avec les bases de données classiques

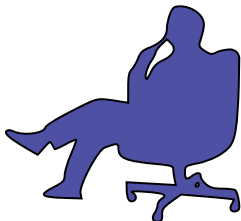
Trois différences principales :

- L'information ne peut être obtenue que par des **vues** (**Local As View**).
- Information **incomplète** et **imprécise**.
- Types **imbriqués**.

Trois sources de complexité !

Direction actuelle de recherche : ensembles **Magic**, **Bucket**, **MiniCon**...

Perspectives



- **k premiers** résultats probabilistes.
- Un cadre d'apprentissage adapté à une annotation imparfaite, qui évite l'**overfitting** (description de longueur minimale?).
- **Déduplication**, identification des **coréférences** malgré des informations légèrement différentes.
- **Corroboration** d'informations entre sources.
- Analyse sémantique élaborée, basée sur une **compréhension** du texte en **langage naturel**.

Autres travaux (1/2)

Similarité entre graphes [SIAM Review 2004, Springer 2008]
(avec Vincent Blondel, UCL); application à
l'**extraction de synonymes** à partir du graphe d'un
dictionnaire.

Identification de la frontière d'un site Web [ICWE 2005] (avec
Serge Abiteboul).

Traduction automatique [MT Summit 2005, XML Conf. 2005]
(avec Jean Senellart, **SYSTRAN**); traitement
complexe de document XML, reconnaissance et
traduction d'entités, transformation de graphes de
relations linguistiques en arbres syntaxiques. . .

Autres travaux (2/2)

Intégration et entrepôts de données pour sociologues [ICWI 2005]

(avec François-Xavier Dudouet, Ioana Manolescu, Benjamin Nguyen); extraction de **réseaux sociaux** à partir de données du Web et de listes de discussion.

Découverte de nœuds similaires dans un graphe [AAAI 2007]

(avec Yann Ollivier, ENS Lyon); application aux **articles connexes** dans Wikipédia.

Prédiction de PageRank (avec Michalis Vazirgiannis, Université d'Athènes).

Merçi.



Vincent D. Blondel, Anahí Gajardo, Maureen Heymans, Pierre Senellart et Paul Van Dooren.

A measure of similarity between graph vertices:
applications to synonym extraction et Web searching.
SIAM Review, 46(4):647–666, 2004.






Pierre Senellart.

Identifying Websites with flow simulation.
Proc. ICWE, Sydney, Australie, juillet 2005.



Mats Attnäs, Pierre Senellart et Jean Senellart.

Integration of SYSTRAN MT systems in an open workflow.
Proc. MT Summit, Phuket, Thaïlande, septembre 2005.

-  François-Xavier Dudouet, Ioana Manolescu, Benjamin Nguyen et Pierre Senellart.
XML warehousing meets sociology.
Proc. IADIS ICWI, Lisbonne, Portugal, octobre 2005.
-  Pierre Senellart et Jean Senellart.
SYSTRAN Translation Stylesheets: Machine translation driven by XSLT.
Proc. XML Conference & Exposition, Atlanta, USA, novembre 2005.
-  Serge Abiteboul et Pierre Senellart.
Querying and updating probabilistic information in XML.
Proc. EDBT, Munich, Allemagne, mars 2006.



Pierre Senellart et Serge Abiteboul.

On the complexity of managing probabilistic XML data.

Proc. PODS, Pékin, Chine, juin 2007.



Yann Ollivier et Pierre Senellart.

Finding related pages using Green measures: An illustration with Wikipedia.

Proc. AAAI, Vancouver, Canada, juillet 2007.



Pierre Senellart et Vincent D. Blondel.

Automatic discovery of similar words.

Michael W. Berry et Malu Castellanos, éditeurs, *Survey of Text Mining: Clustering, Classification et Retrieval.*

Springer-Verlag, janvier 2008.



Pierre Senellart, Avin Mittal, Daniel Muschick, Rémi Gilleron et Marc Tommasi.

Automatic wrapper induction from hidden-Web sources with domain knowledge.

Soumis pour publication à WWW 2008.



Pierre Senellart et Georg Gottlob.

On the complexity of deriving schema mappings from database instances.

Soumis pour publication à PODS 2008.