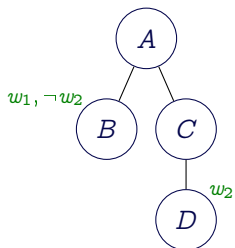


État des recherches

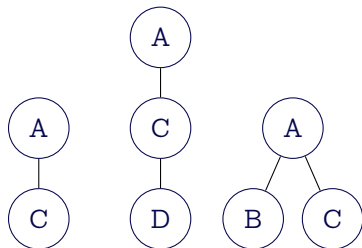
Pierre Senellart



TELECOM ParisTech
27 juin 2008

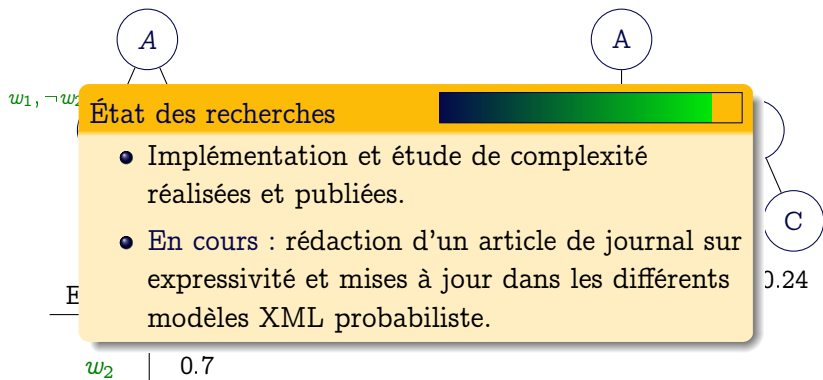


sémantique



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

Event	Prob.
w_1	0.8
w_2	0.7



Correspondance de schémas à partir d'instances

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' prédit

a	a
b	b
c	c
d	d

Correspondance de schémas à partir d'instances

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' prédit

a	a
b	b
d	d

Correspondance de schémas à partir d'instances

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' prédit

a	a
b	b
c	a
d	d

Correspondance de schémas à partir d'instances

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

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

$$R'(c, a)$$

$$R'(g, h)$$

R' prédit

a a

b b

c c

d d

g h

Correspondance de schémas à partir d'instances

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

$$\forall x R(x) \wedge x \neq c \rightarrow R'(x, x)$$
$$\exists y_1 \exists y_2 R'(y_1, y_2) \wedge y_1 = c \wedge y_2 = a$$
$$\exists y_1 \exists y_2 R'(y_1, y_2) \wedge y_1 = g \wedge y_2 = h$$

R' prédit

a	a
b	b
c	c
d	d
g	h

Correspondance de schémas à partir d'instances

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

$$\forall x R(x) \wedge x \neq c \rightarrow R'(x, x)$$
$$\exists y_1 \exists y_2 R'(y_1, y_2) \wedge y_1 = c \wedge y_2 = a$$
$$\exists y_1 \exists y_2 R'(y_1, y_2) \wedge y_1 = g \wedge y_2 = h$$

Coût : 17

R' prédit

a	a
b	b
c	c
d	d
g	h

Correspondance de schémas à partir d'instances

R	R'
a	a a
b	b b

État des recherches

- Étude détaillée de complexité
- En cours : amélioration de certains résultats, rédaction de papier journal
- Possibles prolongements : étude des liens avec la programmation logique inductive

$\exists y_1$

édit
a
b
c c
d d
g h

Collaboration avec Georg Gottlob, University of Oxford.



Extraction d'informations depuis des pages de résultat

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 Bouganin](#) (INRIA Rocquencourt), [Cosmin Cremarencu](#) (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

Collaboration avec l'équipe-projet Mostrare, INRIA Lille.



Extraction d'informations depuis des pages de résultat

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: VOFitter, Bridging Virtual Observatory and Industrial Office Applications

Authors: [Chen-Zhou Cui](#) (1), [Markus Dolensky](#) (2), [Peter Quinn](#) (2), [Yong-heng Zhao](#) (1), [Francoise Genou](#) (3) ((1)NAO China, (2)ESO, (3)CDS)

Comments: Accepted for publication in CHI'AA (9 pages, 2 figures, 185KB)

3. [cs.DS/0512061](#) [abs, ps, pdf, other] :

Title: Matching Subsequences in Trees

Authors: [Philippe Gibb](#), [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](#)), [Cormac Cremers](#) ([INRIA Rocquencourt](#)), [Francis Dang Ngoc](#) ([INRIA Rocquencourt](#)), PRISM - UVSQ,

[Nicolas Dreu](#) ([INRIA Rocquencourt](#)), [Philippe Rucheral](#) ([INRIA Rocquencourt](#)), PRISM - UVSQ)

Subj-class: Cryptography and Security; Databases

Collaboration avec l'équipe-projet Mostrare, INRIA Lille.



Extraction d'informations depuis des pages de résultat

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) 2006

Subj-class: Logic in Computer Science; Cryptography and Security

ACM-class: H.2.7; K.4.4

2. **astro-ph/0512493** [abs, pdf] :

Title: VOFiler, Bridging Virtual Observatory and Industrial Office Applications

Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Françoise Genou (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 Bink, 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 Desprez (1), Guillaume Huet (2), Ingrid Isenhardt (3)

Subj-class: Information Retrieval

5. **cs.CR/0510013** [abs, pdf] :

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: Luc Bouganim (1), Renaud Boudreau (2), Coenraed Cremerence (3), John Fournier (4), François Dang Ngoc (5), Inria Mesquenaux (6), PRISM - UVSQ)

Nicolas Dreu (7), Guillaume Huet (8), Philippe Pucheral (9), Guillaume Huet (10), PRISM - UVSQ)

Subj-class: Cryptography and Security; Databases

- Première annotation imprécise et incomplète grâce à la connaissance du domaine.
- Affinage par généralisation structurelle du document (champs aléatoires conditionnels, appliqués de manière non supervisée !).
- Permet d'obtenir sans intervention humaine un extracteur (*wrapper*) des résultats.

Collaboration avec l'équipe-projet Mostrare, INRIA Lille.



Extraction d'informations depuis des pages de résultat

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) 2006

Subj-class: [Logic in Computer Science](#): Cryptography and Security

ACM-class: H.2.7; K.4.4

2. [astro-ph/0512493](#) [abs, pdf] :

Title: VOFiler, Bridging Virtual Observatory and Industrial Office Applications

Authors: [Chen-Zhou Cui](#) (1), [Markus Dolensky](#) (2), [Peter Quinn](#) (2), [Yong-heng Zhao](#) (1), [Francoise Genou](#) (3) ((1)NAO China, (2)ESO, (3) CDS)

Comments: Accepted for publication in CHI'AA (9 pages, 2 figures, 185KB)

3. [cs.DS/0512061](#) [abs, ps, pdf, other] :

État des recherches

- Premières expériences réalisées, résultats satisfaisants, article en cours de revue
- À venir : expériences sur d'autres domaines, amélioration du cadre d'apprentissage pour obtenir de meilleurs résultats
- Filtrage par **generalisation structurelle** du document (champs aléatoires conditionnels, appliqués de manière **non supervisée**!).
- Permet d'obtenir **sans intervention humaine** un extracteur (*wrapper*) des résultats.

Collaboration avec l'équipe-projet Mostrare, INRIA Lille.

Prédiction de PageRank

- Le score (PageRank) d'une page Web **évolue** au cours du temps.
- **Coûteux** de parcourir tout le Web afin de mettre à jour ces scores.
- Possibilité de **prévoir** ces évolutions, en identifiant des tendances récurrentes ?

- Le score (PageRank) d'une page Web **évolue** au cours du temps

État des recherches

- Bon résultats avec des modèles de Markov cachés
- Poster publié, papier en cours de revue

- Quels services (du Web caché) interroger pour répondre à une requête d'un utilisateur ?
- Service, Requête : règle Datalog
- Différences avec interrogation classique de base de données :
 - L'information ne peut être obtenue que par des **vues** (**Local As View**)
 - Restriction sur les accès à ces vues (**binding patterns**)
 - Information **incomplète** et **imprécise**
 - Types **imbriqués**

- Quels services (du Web caché) interroger pour répondre à une requête ?

État des recherches

- Modèle défini
- Pas de solution évidente
- Beaucoup de littérature sur des thèmes voisins, mais ne répondant pas exactement au problème
- En cours : exploration du côté de Magic, Minicon, Inverse-Rules...

ées :

Estimer la date de dernière modification d'une page :

- Timestamp et ETag HTTP
- Timestamp dans le contenu (max des timestamps ?)
- Comparaison avec une ancienne version, sans tenir compte des différences non significatives (hachage du texte uniquement, shingling, distance d'édition...)
- Méta-information sémantique (RSS, sitemaps, timestamp de documents PDF ou Word...)

Estimer la date de dernière modification d'une page :

- T État des recherches
- T ● Différentes sources d'informations répertoriées
- C ● À venir : expérimentation systématique de la ripte
d qualité de ces sources, pour définir une
u stratégie de datation
- M ● Gros crawls du Web à disposition (European ript de
d Archive)

Comment utiliser la **redondance** des faits déclarés par différentes sources (p. ex., sur le Web) pour estimer la **confiance** en ces faits ?

- Graphe bipartite source/faits
- Certaines sources nient des faits
- Certains faits sont incompatibles (dépendances fonctionnelles)
- etc.

Comment utiliser la **redondance** des faits déclarés par différentes sources (ex: $\{M, H, \dots\}$) ?

État des recherches

- Modèle et algorithmes à peu près élaborés dans certains cas (PageRank en biparti, OPIC avec cash négatif)
- En cours : implémentation, expérimentations
- À venir : généralisations, dépendances fonctionnelles souples, etc.

- DataRing : base de données semi-structurées auto-administrée
- Problème spécifique : gestion d'incertitude (existence d'informations, mesures imprécises, etc.)

- DataRing : base de données semi-structurées

État des recherches

- À faire : à peu près tout ! Basé sur XML probabiliste ?