

SXPath - Extending XPath towards Spatial Querying on Web Documents

Ermelinda Oro¹ Massimo Ruffolo¹ Steffen Staab²

¹Institute of High Performance Computing and Networking of CNR (ICAR-CNR)
University of Calabria, Italy

²Institute for Computer Science, University of Koblenz, Koblenz, Germany

VLDB 2011



Outline

1 Introduction

- Motivations
- State of the Art
- SXPath Language

2 SXPath

- Spatial Data Model
- Syntax and Semantics
- Complexity Issues
- Implementation Issues and Experiments

3 Conclusions and Future Work



Outline

1 Introduction

- Motivations
- State of the Art
- SXPath Language

2 SXPath

- Spatial Data Model
- Syntax and Semantics
- Complexity Issues
- Implementation Issues and Experiments

3 Conclusions and Future Work

Motivations

- Users need to access the Web and capture information in many application fields (e.g. business, competitive and military intelligence; content, document and knowledge management)
- Web pages are human oriented. The spatial arrangement of content items in Web pages produces visual cues that help human readers to make sense of document contents
- Well founded and known query formalisms, such as XPath and XQuery, do not consider spatial arrangements in querying Web pages

Presentation-Oriented Documents

sign in with join CNET

[Reviews](#) [News](#) [Downloads](#) [Video](#) [How To](#) [Search](#)

Parallels Desktop 6 per Mac [dai un'occhiata](#)

SONY PRODUCT REVIEW RESULTS FOR "camera" [COMPILE SELECTED](#)

Sony Cyber-shot DSC-HX9V (Black)
EDITORIAL RATING: [USER RATING:](#) [out of 8 reviews](#)

Feature junkies in search of a compact megazoom should get the Sony Cyber-shot DSC-HX9V.

Canon PowerShot S95
EDITORIAL RATING: [USER RATING:](#) [out of 23 reviews](#)

Though well-made to perform better, dSLR shooters looking for a sleek compact camera will find the Canon PowerShot S95's top-flight photos and a full manual feature set worth the ticket of its compact size.

Canon PowerShot ELPH 300 HS (black)
EDITORIAL RATING: [USER RATING:](#) [out of 2 reviews](#)

If you're looking for a nice, simple ultracompact for use indoors and out, you'll want to strongly consider the Canon PowerShot ELPH 300 HS.

PRODUCT REVIEWS (2,486)

Camera (Non-digital) Accessories (5,005)
Cameras (Non-digital) (274)
Digital Camera Accessories (4,832)
Digital Camera (5,005)

See all matching products

NEWS, PHOTOS & VIDEOS (17,697)

Post 7 Days (93)
Post 30 Days (334)
Post Year (3,666)

SOFTWARE DOWNLOADS (16,790)

Mac (196)
Mobile (11,774)
Webware (31)
Windows (4,191)

MediaWorld
COMPRO ON-LINE

800-882288
Acquista via telefono

CATALOGO

NET-PRINT stampanti, foto

NET-MUSIC audiocassette, musica

NET-MOVIE videocassette, film

NET-GIFT biglietti, regali

NET-BOOK new laptop, netbook

E-VIDEOPLAYING TV
televisori, video

Sal refresher: [mediaworld.refresher](#)

RSS: [RSS](#)

I tuoi dati: [Supporto clienti](#) | [Contatti](#)

Carrello: 0

CERCA: Tutte le categorie **Vali**

[Nuova ricerca](#) [Risultati - LogIn](#)

Totale articoli: **31**

I prezzi di vendita sono comprensivi di IVA, contributo PAEE.

NIKON D3000 KIT 18-55 VR
Fotocamera Reflex digitale - Sensore da 10 Megapixel
LCD da 3" - Full TTL - Sist. SD/SDHC - Uscita video
Obiettivo 18-55 MM - Fiss. 4.5-5.6 g.
Garanzia ufficiale 3 anni NIKE Italia

399,00

Scheda di memoria da 4GB contenuta nella confezione
Consegna gratuita in tutta Italia

Vel a pagina: **1** [2](#) [3](#) [4](#) [5](#)

SONY DSC-HX30V KIT 18-55
Fotocamera Reflex digitale - Sensore da 14 Megapixel
LCD da 2,7" - Live View - Display anamorfico
3x zoom ottico - Sist. SD/SDHC - Uscita video
Obiettivo 18-55 MM - Uscita HDMI - Fiss. 4.5-5.6 g.
Garanzia ufficiale Sony Italia

399,00

NON DISPONIBILE - [Avvisi](#) [Qualità](#) [Copia](#) [Scegli](#)

SOLD OUT

399,00

P

Prezzo di mercato €449,00
Iva esclusa €36,00 pari al 8,11,13% di acconto.

[Aggiungi al carrello](#)

New User? Register | Sign In | Help

Upload to Photos 4

Mail | My Y! | Yahoo!

YAHOO! SHOPPING

Search

Search Web

Home Clothing Electronics Computers Home & Garden Shopping Insider More

Shop for: cameras reflex

SEARCH In All departments More

Shopping > Cameras > Digital Cameras & Accessories > Digital Cameras > cameras

1 - 18 out of 2,384 results for cameras (About)

Digital Cameras

Narrow Results (camera)

▼ Price

\$0.00 - \$600.00
\$600.00 - \$900.00
\$900.00 - \$1,000.00
\$1,000.00 - \$1,400.00
Above \$1,400.00

From \$ To \$ Refine

Results per page: 15 | 30 | 45 Sort by: Top Results Show grid view ▾

Canon EOS Rebel T2i Black SLR Digital Camera... \$674.95 - \$1,199.99 Compare prices

 The Canon EOS Rebel T2i brings professional EOS features into an easy-to-use, compact digital SLR that's a joy to use.

Pros: 18mpf easy to use but function rich
Cons: Battery charges extremely slowly

Compare  Write a review

Canon Powershot ELPH 300 HS Black Digital Camera \$220.00 - \$249.97 Compare prices

 Canon Powershot 300 HS digital cameras smooth and lustrous eye-catching design. Made for your hands, your choices.

Compare  Write a review

Web Images Videos Maps News Shopping Gmail more .

Google

camera

About 8,106,000 results (0.45 seconds)

Set your location Advanced search

Sort by: Relevance

Canon PowerShot SX230 HS 12 Megapixel Digital Camera - Black
Canon - 12.1 megapixel - 14x optical zoom - Point & Shoot
Canon EOS 50D SLR - Canon's HS SYSTEM with a 12.1 Megapixel CMOS and DIGIC 4 image Processor improves shooting in low-light situations ...
★★★★★ 30 reviews - Add to Shopping List
#5 in Point & Shoot Canon Digital Cameras

Canon EOS 600D Digital SLR Camera with Canon EF-S 18-135mm IS lens
Canon - 18 megapixel - Standard SD - BDH - 500x - 7.5 x optical zoom - ISO 12800 - 1080p Full HD Video - 18 MP CMOS Sensor
With the EOS 600D DSLR, Canon gives the photo enthusiast a powerful tool featuring better image quality, more advanced features ...
★★★★★ 389 reviews - Add to Shopping List
#5 in SLR Canon Digital Cameras

Canon EOS 70D Digital SLR Camera with Canon EF 28-135mm IS lens
Canon - 18 megapixel - Compact Flash - 5 x optical zoom - ISO 12800 - Pop-up Flash - 28.8 optical - Optical Viewfinder - SLR
The EOS 70D is designed to entertain every facet of the photographic process, from still image capture to video. The EOS 70D features a host of new ...
★★★★★ 473 reviews - Add to Shopping List

\$515 from 26 stores Compare prices

\$1,000 from 63 stores Compare prices

\$1,630 from 75 stores Compare prices

Presentation-Oriented Documents

HTML DOM allows only site-centric extraction

A Web Page

Document Object Model

last.fm Musica Video Radio Eventi Classifiche Musica

Trova musica su Last.fm

Cerca in catalogo: Inserisci artista o tag Cerca

Tutti

- acoustic
- ambient
- blues
- classical
- country
- electronic
- emo
- folk
- gothic
- hardcore

Musica più ascoltata su Last.fm

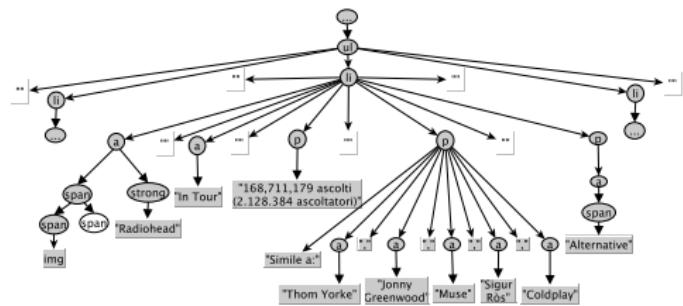
Più ascoltata | Del momento | Più ascoltata in Germania

Coldplay
114,516,462 ascolti (2,300,610 ascoltatori)
Smile a: Keane, Travis, The Killers, Snow Patrol, Oasis
- rock

IN TOUR


Radiohead
168,711,179 ascolti (2,128,384 ascoltatori)
Smile a: Thom Yorke, Jonny Greenwood, Muse, Sigur Rós, Coldplay
- alternative

IN TOUR

Spatial arrangements are rarely explicit and frequently hidden in complex nestings of layout elements corresponding to intricate tree structures that are conceptually difficult to query

Outline

1 Introduction

- Motivations
- State of the Art
- SXPath Language

2 SXPath

- Spatial Data Model
- Syntax and Semantics
- Complexity Issues
- Implementation Issues and Experiments

3 Conclusions and Future Work

State of the Art

- Web Query language
 - *XPath 1.0* and *XQuery 1.0* represent well founded and known web query languages having very intuitive navigational features, but the intricate DOM structure makes difficult to pose queries
- Visual languages
 - *Spatial Graph Grammars* [Kong et al.] are quite complex in term of both usability and efficiency
 - *Algebras* for creating and querying multimedia interactive presentations (e.g. ppt) [Subrahmanian et al.] require database for multimedia presentation should be created for the whole Web
- Web wrapper induction exploiting visual interface [Gottlob et al.] [Sahuguet et al.]
 - generate XPath location paths of DOM nodes
 - can benefit from using Spatial XPath

Outline

1 Introduction

- Motivations
- State of the Art
- SXPath Language

2 SXPath

- Spatial Data Model
- Syntax and Semantics
- Complexity Issues
- Implementation Issues and Experiments

3 Conclusions and Future Work

Extending XPath towards Spatial Querying

- As extension of XPath 1.0, *Spatial XPath* (SXPath):
 - adopts the intuitive path notation: $/axis::nodetest [pred_1] *$
 - adds new *spatial axes* and new *spatial position functions*
 - has a natural semantics that enables spatial querying
 - maintains polynomial time combined complexity
- Advantages:
 - it is easy to learn and easier to use than pure XPath on Web pages
 - it is more tolerant to modifications of the internal structure of Web pages
 - it enables users to spatial query Web documents on the base of what they see on the document
 - it is capable to provide benefits to some current Web contents manipulation and wrapper learning approaches

Presentation-Oriented Documents

A Web Page from the lastfm Web site (<http://www.lastfm.it/>)

Acquiring a music band profile: *A music band photo that has at least its descriptive information*

The screenshot shows the last.fm homepage with a red header bar. The header includes the last.fm logo, navigation links for Musica, Video, Radio, Eventi, and Classifiche, a search bar, and links for Accesso and Registrati. Below the header, a sidebar on the left lists various music genres: Tutti, acoustic, ambient, blues, classical, country, electronic, emo, folk, gothic, hardcore, hip hop, and indie. The main content area features a large box titled "Musica più ascoltata su Last.fm" with sub-sections for "Più ascoltata" (Coldplay, 114,516,462 ascolti), "Del momento" (Radiohead, 168,711,179 ascolti), and "Più ascoltata in Germania" (Coldplay). Each section includes a photo of the band, a "IN TOUR" badge, a list of similar artists, and a genre tag (rock for Coldplay, alternative for Radiohead). A "Tra i tuoi consigli" section on the right encourages users to sign up for personalized music recommendations. The footer of the page includes the CAR logo.

Example 1



Radiohead

168,711,179 ascolti (2,128,384 ascoltatori)

Simile a: [Thom Yorke](#), [Jonny Greenwood](#), [Muse](#),
[Sigur Rós](#), [Coldplay](#)

• alternative

Exploiting XPath

```
for $li in document
("last-fm.htm")
(1.1) //div[@id='content'] //ul/li
return
<music-band>
(1.2) <name>
      {$ li / a / strong / text()}
    </name>
...
</music-band>
```

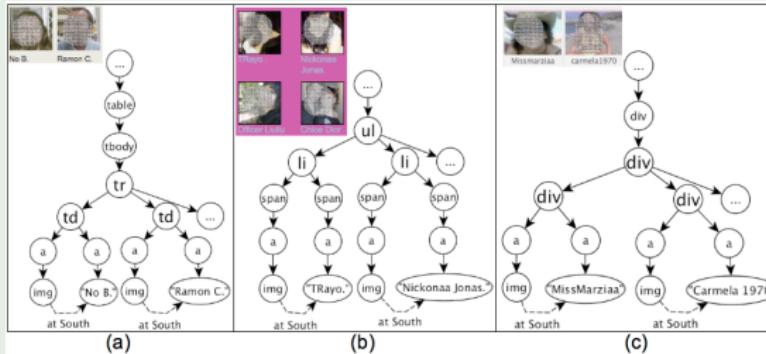
Exploiting SXPath

```
for $li in document
("last-fm.htm")
(2.1) / CD::img [N|S::img]
return
<music-band>
(2.2) <name>
      {$img/ E::text [N,1]}
    </name>
...
</music-band>
```

Example 2

Acquiring friend lists from different social networks pages represented as couples `<photo, name>`.

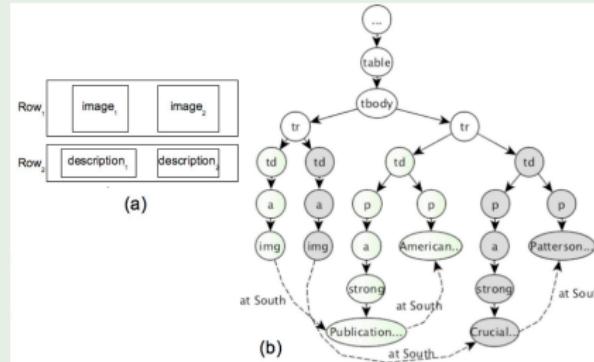
Friend lists from different social networks pages (a) Bebo (b) Care (c) Netlog.



```
for $img in document ("http://www.bebo.com/friendlist.html")
(3.1) //img[ N|S|E|W::img ]
  return
    <friend>
(3.2) <photo> {$img} </photo>
(3.3) <name> { $img/ S :: text() [N,1] } </name>
  </friend>
```

Example 2

- A single data record can be split in different sub-trees
- Wrapper induction techniques like DEPTA [Zhai et al.] recognize data records when they are encoded in the DOM as consecutive similar subtrees



```

for $img in document ("http://www.bebo.com/friendlist.html")
(3.1) //img[ N|S|E|W::img ]
  return
    <friend>
(3.2) <photo> {$img} </photo>
(3.3) <name> { $img/ S :: text() [N,1] } </name>
    </friend>
  
```

Outline

1 Introduction

- Motivations
- State of the Art
- SXPath Language

2 SXPath

- **Spatial Data Model**
- Syntax and Semantics
- Complexity Issues
- Implementation Issues and Experiments

3 Conclusions and Future Work

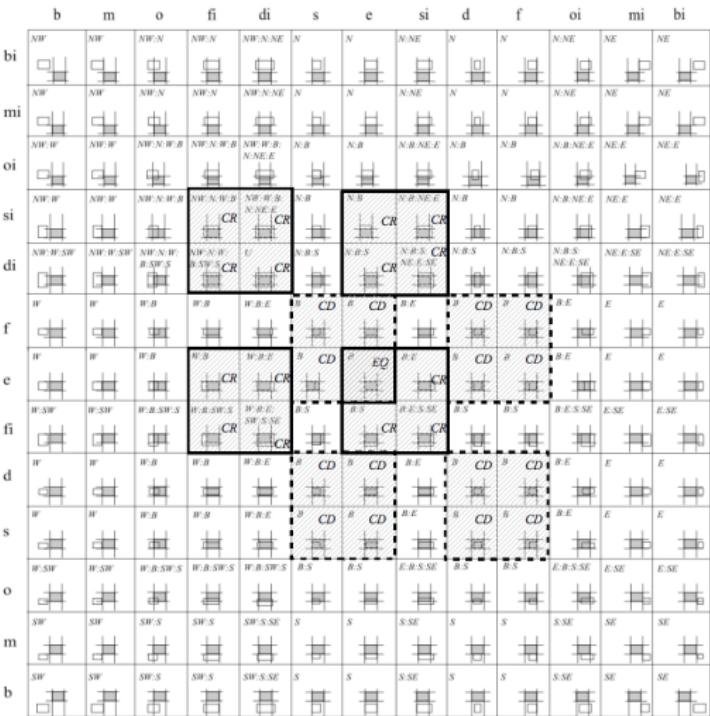
Spatial Data Model

- The Document Object Model (DOM) is the internal representation of markup languages (XML, HTML)
- The tree-based structures of XML are often not convenient and not expressive enough in order to represent spatial arrangements
- The spatial arrangements are rarely explicit and frequently hidden into intricate tree structures that are conceptually difficult to query

Spatial Relations among Nodes

- The *Rectangular Algebra* (RA) [Balbiani et al.] extends Allen's temporal interval algebra (IA) to the 2-dimensional case
- RA is a very fine-grained and expressive model that allows the computations of spatial relations as well as algebraic optimizations
- RA holds many important properties (e.g. invertibility) that allows for optimized query evaluation

Relation	Symbol	Meaning	Inverse
before	b	$s_1 \xrightarrow{\text{---}} s$	bi
meets	m	$s_1 \xrightarrow{\text{---}} s$	mi
overlaps	o	$s_1 \xrightarrow{\text{---}} s$	oi
starts	s	$s_1 \xrightarrow{\text{---}} s$	si
during	d	$s_1 \xrightarrow{\text{---}} s$	di
finish	f	$s_1 \xrightarrow{\text{---}} s$	fi
equals	e	$s_1 \xrightarrow{\text{---}} s$	e

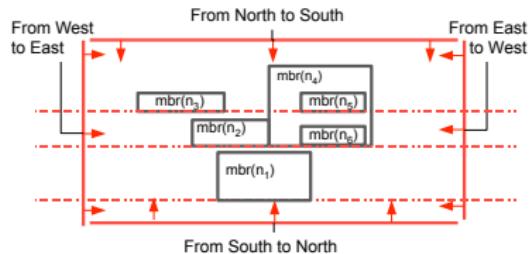
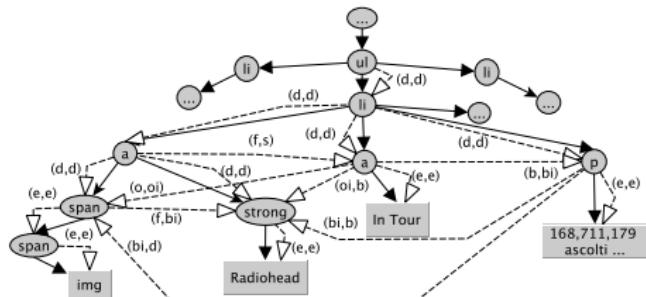


Spatial DOM (SDOM)



The SDOM extends the Document Object Model (DOM) by:

- RA relations existing between pairs of nodes visualized on screen
- spatial orders among nodes



$$n_1 \leqslant n_2 = \uparrow n_4 = \uparrow n_6 \leqslant n_3 = \uparrow n_5$$

The Spatial DOM (SDOM)

Definition

SDOM is a node labeled sibling ordered tree enriched by RA relations

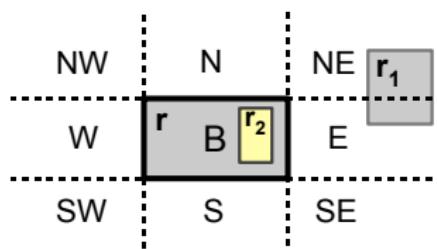
$$SDOM = \langle V, R_{\Downarrow}, R_{\Rightarrow}, A, f_s \rangle$$

where:

- V is the set of labeled DOM nodes. $V = V_v \cup V_{nv}$
- R_{\Downarrow} is the *firstchild* relation
- R_{\Rightarrow} is the *nextsibling* relation
- $A \subseteq V_v \times V_v$
- Let R_{rec} be the set of RA relations $f_s : A \rightarrow R_{rec}$

Qualitative Spatial Models

Rectangular cardinal relations



Topological relations, inspired by the Region Connection Calculus model:

- contained (CD)
- container (CR)
- equivalent (EQ)

Example

- $r \text{ E:NE } r_1$
- $r \text{ B } r_2$

Example

- $r \text{ CD } r_2$
- $r_2 \text{ CR } r$

Spatial Navigation Axes

- As in XPath, SXPath primitives for navigating the SDOM are called axes
- Axes are interpreted binary relations $\chi \subseteq V \times V$. Let $self := \{(u, u) | u \in V\}$ be the reflexive axis, remaining SXPath axes are partitioned in two sets: Δ_t and Δ_s
 - $\Delta_t = \{self, child, parent, descendant, descendant-or-self, ancestor, ancestor-or-self, following-sibling, preceding-sibling, following, preceding\}$ contains traditional XPath 1.0 axes
 - Δ_s is the set of novel spatial axes expressed by: basic and disjunctive RCRs and topological relations that are more intuitive than RA relations

Spatial Navigation Axes

Definition

SXPath spatial axes are interpreted binary relations

$\chi_s \subseteq V_v \times V_v$ of the following form

$\chi_s = \{\langle u, w \rangle | u, w \in V_v \wedge u \rho w \wedge \rho \in \mu(R)\}$. Where R is the RC or Topological Relation that names the spatial axis and μ is the mapping function

	b	m	o	fi	di	s	e	si	
bi									
mi									
oi									
si									
di									

Outline

1 Introduction

- Motivations
- State of the Art
- SXPath Language

2 SXPath

- Spatial Data Model
- **Syntax and Semantics**
- Complexity Issues
- Implementation Issues and Experiments

3 Conclusions and Future Work

Syntax

- SXPath expressions have the same structure as the ones in XPath
 - Location paths are sequences of location steps separated by the navigation operator "/".
 - A locstep is `axis :: nodetest [pred1]... [predn]`
- We enrich XPath 1.0 by
 - The new set of *spatial axes*
 - *Spatial position functions*
- Specific subsets of the language with attractive properties have been characterized for XPath 1.0 [4, 6]
 - *Core XPath* ⇒ *Core SXPath*
 - *Wadler Fragment(WF)* ⇒ *Spatial WF*

Semantics

- The main structural feature of SXPath are *expressions*, that return a value from one of the following four types: *node set*, *number*, *string*, or *Boolean*
- Every expression evaluates relative to a *context*, concept introduced by Wadler

Definition (Context)

The *context* is the following 12-tuple:

$$\vec{c} = \langle n, p_{\leqslant_{\text{doc}}}, s_{\leqslant_{\text{doc}}}, p_{\leqslant_{\uparrow}}, s_{\leqslant_{\uparrow}}, p_{\leqslant_{\rightarrow}}, s_{\leqslant_{\rightarrow}}, p_{\leqslant_{\downarrow}}, s_{\leqslant_{\downarrow}}, p_{\leqslant_{\leftarrow}}, s_{\leqslant_{\leftarrow}}, p_{\leqslant_{\text{f}}} \rangle$$

where:

- n is a *context node*
- p_{\leqslant_z} are the *context positions* w.r.t. orders
- s_{\leqslant_z} are the *context sizes*

Semantics

Definition (Location path semantics)

Let π, π_1, π_2 be location paths, let $locstep$ be a location step over an axis χ , let $bexpr$ be a boolean expression and let n be a context node, $P: LocationPath \rightarrow node \rightarrow nodeset$ is defined as follows:

$$P[\![/\pi]\!](n) := P[\!]\pi\!](root)$$

$$P[\!\pi_1/\pi_2]\!](n) := \{n_2 | n_1 \in P[\!\pi_1]\!](n) \wedge n_2 \in P[\!\pi_2]\!](n_1)\}$$

$$P[\!\pi_1|\pi_2]\!](n) := P[\!\pi_1]\!](n) \cup P[\!\pi_2]\!](n)$$

$$P[\!axis :: t]\!](n) := \{n' | \text{[axis]}(n, n')\} \cap T(t)$$

$$P[\!locstep[bexpr]\!](n) := \{n' | \vec{W} = P[\!locstep]\!](n) \wedge n' \in \vec{W} \wedge \varepsilon[\!bexpr]\!(\vec{c}_{n'}) = \text{true} \wedge \vec{c}_{n'} := \langle n', \text{idx}_{\chi}(n', \vec{W}), |\vec{W}|, \text{pidx}_{\leqslant \uparrow}(n', \vec{W}), \text{plast}_{\leqslant \uparrow}(\vec{W}), \text{pidx}_{\leqslant \rightarrow}(n', \vec{W}), \text{plast}_{\leqslant \rightarrow}(\vec{W}), \text{pidx}_{\leqslant \downarrow}(n', \vec{W}), \text{plast}_{\leqslant \downarrow}(\vec{W}), \text{pidx}_{\leqslant \leftarrow}(n', \vec{W}), \text{plast}_{\leqslant \leftarrow}(\vec{W}), \text{pidx}_{\leqslant t}(n', \vec{W}) \rangle\}$$

The semantics of spatial axis is given in terms of spatial relations among nodes
[spatialAxis] := $\{(n, n') | mbr(n) \rho mbr(n') \wedge \rho = \mu(\text{spatialAxis})\}$

Semantics

Definition (Semantics of SXPath)

$$\varepsilon : \text{SXPathExpression} \rightarrow \mathbf{C} \rightarrow \text{SXPathType}$$

$$\varepsilon[\pi](\vec{c}) := P[\pi](n)$$

$$\varepsilon[\text{position}()](\vec{c}) := p_{<_{\text{doc}}}$$

$$\varepsilon[\text{posFromN}()](\vec{c}) := p_{\leqslant \downarrow}$$

$$\varepsilon[\text{posFromS}()](\vec{c}) := p_{\leqslant \uparrow}$$

$$\varepsilon[\text{posFromW}()](\vec{c}) := p_{\leqslant \rightarrow}$$

$$\varepsilon[\text{posFromE}()](\vec{c}) := p_{\leqslant \leftarrow}$$

$$\varepsilon[\text{posSpatialNesting}()](\vec{c}) := p_t$$

$$\varepsilon[\text{Op}(e_1, \dots, e_m)](\vec{c}) := F[\text{Op}](\varepsilon[e_1](\vec{c}), \dots, \varepsilon[e_m](\vec{c}))$$

$$\varepsilon[\text{last}()](\vec{c}) := s_{<_{\text{doc}}}$$

$$\varepsilon[\text{lastFromN}()](\vec{c}) := s_{\leqslant \downarrow}$$

$$\varepsilon[\text{lastFromS}()](\vec{c}) := s_{\leqslant \uparrow}$$

$$\varepsilon[\text{lastFromW}()](\vec{c}) := s_{\leqslant \rightarrow}$$

$$\varepsilon[\text{lastFromE}()](\vec{c}) := s_{\leqslant \leftarrow}$$

$$\begin{aligned} F[\text{RelOp: num} \times \text{num} \rightarrow \text{bool}](i_1, i_2) &::= i_1 \text{ RelOp } i_2 \\ F[\text{constant number } i : \rightarrow \text{num}](i) &::= i \end{aligned}$$

...

Outline

1 Introduction

- Motivations
- State of the Art
- SXPath Language

2 SXPath

- Spatial Data Model
- Syntax and Semantics
- **Complexity Issues**
- Implementation Issues and Experiments

3 Conclusions and Future Work

Core SXPath Complexity

Theorem (Core SXPath Combined Complexity)

*Core SXPath queries can be evaluated in time $O(|D|^2 * |Q|)$ where $|D|$ is the size of the XML document, and $|Q|$ is the size of the query Q*

- **Proof Sketch** There are $O(|V_v|^2)$ many spatial relations to be considered in addition to the $O(|V|)$ many relations of the DOM incurring a higher polynomial worst case complexity

SWF and Full SXPath Complexity

Theorem (Spatial WF Combined Complexity)

time $O(\max(|D|^3 * |Q|, |D|^2 * |Q|^2))$ and space $O(|D|^2 * |Q|^2)$, where D is the XML document, and Q is a SWF query.

Theorem (Full XPath Combined Complexity)

time $O(|D|^4 * |Q|^2)$ and space $O(|D|^2 * |Q|^2)$, where D is the XML document, and Q is a Full SXPath query.

- In order to obtain a polynomial-time combined complexity bound for SXPath query evaluation we use dynamic programming adopting the *Context-Value Table* (CV-Table) principle introduced by Gottlob et al.
- Position and size are computed on demand, we compute all spatial position functions in a loop for all pairs previous\current nodes
- Full SXPath computational costs are dominated by String Operations belonging to XPath 1.0
- In SWF the computation of spatial ordering generates a higher polynomial worst case than XPath 1.0



Complexity Results

Comparison between complexity bound of SXPath and XPath 1.0 for a XML document D and a query Q

	XPath 1.0		SXPath	
Space	Core	$O(D * Q)$	Spatial	$O(D ^2 * Q)$
Time		$O(D * Q)$	Core	$O(D ^2 * Q)$
Space	EWF	$O(D * Q ^2)$	SWF	$O(D ^2 * Q ^2)$
Time		$O(D ^2 * Q ^2)$		$O(\max(D ^3 * Q , D ^2 * Q ^2))$
Space	Full	$O(D ^2 * Q ^2)$	Full	$O(D ^2 * Q ^2)$
Time	Xpath 1.0	$O(D ^4 * Q ^2)$	SXPath	$O(D ^4 * Q ^2)$

Outline

1 Introduction

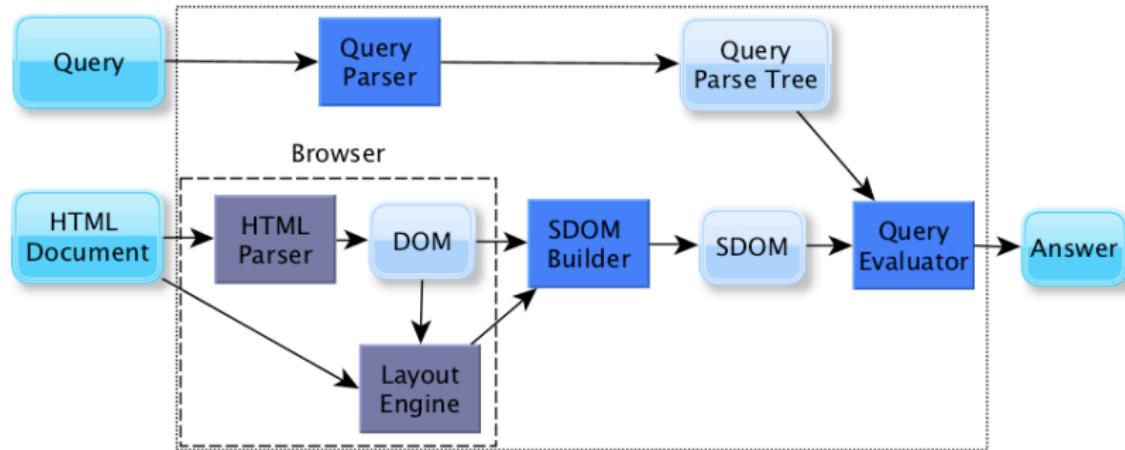
- Motivations
- State of the Art
- SXPath Language

2 SXPath

- Spatial Data Model
- Syntax and Semantics
- Complexity Issues
- **Implementation Issues and Experiments**

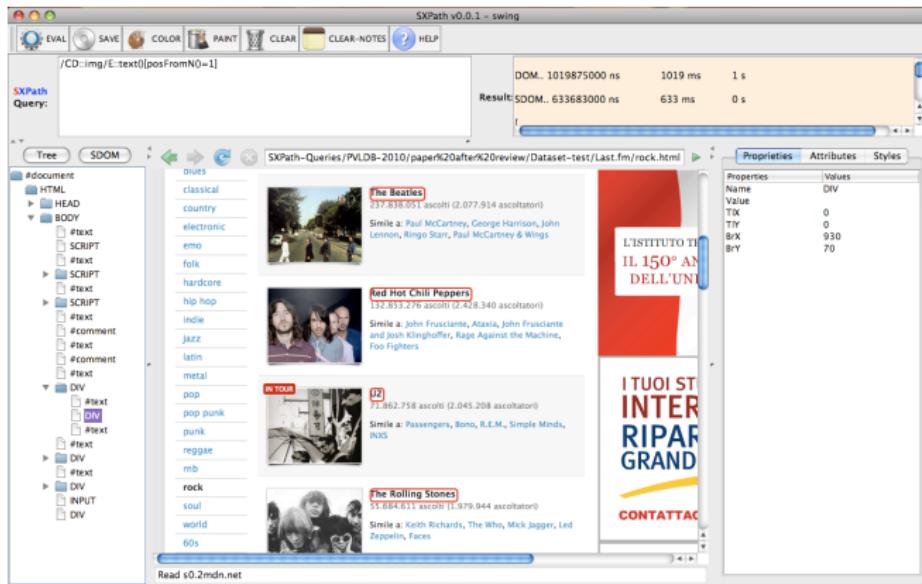
3 Conclusions and Future Work

The SXPath System



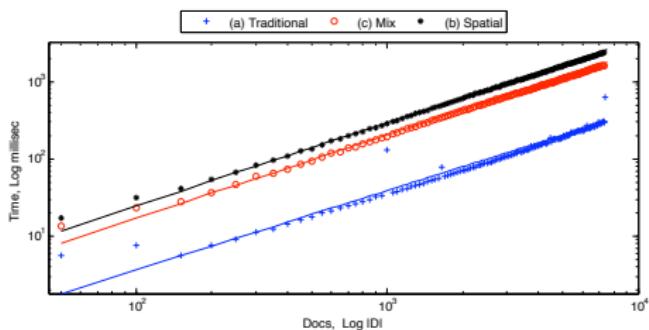
The XPath System

GUI that supports Spatial Querying

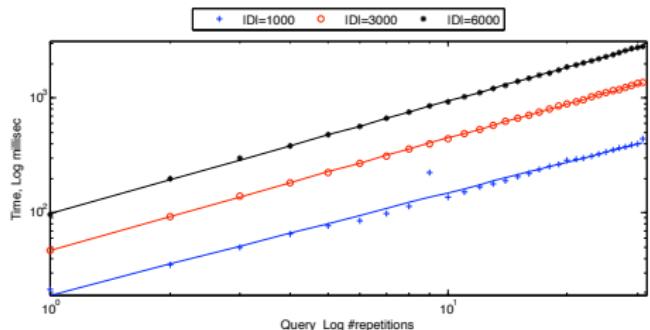


Results of Experiments

Data Efficiency of SXPath Query Evaluation



Query Efficiency of SXPath Query Evaluation



The curves grows linear on log-log scale indicating the polynomial growth

Results of Experiments

Evaluation of the Effort Needed for Learning and Applying SXPath

- We have defined the user task “identify product data records and extract product names and prices” from the Web site www.bol.de
- We have asked users to learn the SXPath language and complete the task by writing a sound and complete SXPath query looking only at the visualized Web page
- We have asked users to answer a questionnaire based on the seven-item Likert scale: very easy/satisfactory (3) ... very difficult/unsatisfactory (-3)

#user	Time (min)	Easiness/ Difficulty	Satisfaction/ Unsatisfaction	#attempts	
				name	price
1	75	2	0	7	6
2	45	3	2	4	2
3	65	1	1	5	4
4	40	2	1	2	3
5	50	3	2	4	4
6	30	3	3	2	1
7	125	-1	-1	9	8
8	50	2	1	3	4
9	35	3	2	2	2
10	55	2	1	5	2
Average	57	2	1.2	4.3	3.6
σ	26	1.18	1.1	2.2	2

Results of Experiments

Usability Evaluation of SXPath on Deep Web Pages

We have asked users to perform the extraction task “identify product data records and extract product names and prices” for each Web site in the dataset

- only by looking at the displayed Web pages by using at the most 5 attempts
- looking at both visualized Web pages and internal page structures (i.e. DOM and SDOM), by using at the most 10 minutes
- by applying the same location path for different Web sites in the dataset having the same visual pattern. We have observed that it is possible to use the same sound and complete spatial location path for Web sites having the same visual pattern. Instead, different XPath location paths are needed

Considering a set of Deep Web Sites	Querying Without DOM/SDOM						Querying With DOM/SDOM					
	SXPath			XPath			SXPath			Abs. XPath		
	Cr.	Wr.	Att.	Cr.	Wr.	Att.	Att.	Steps	Att.	Steps	Att.	Steps
Average			2			5	2.7	5.3	4.2	18.9	4	6.6
Total	2535	27.3/6		2506	3459.5/35							
Recall	100%			99%								
Precision	99%			42%								

Conclusions and Future Work

- We have extended XPath to include spatial navigation into the query mechanism
- The SDOM extends DOM for describing relationships between data entities
- SXPath query language is a stepping stone for future work on extracting information from presentation-oriented documents. It could be used and extended for
 - querying other *presentation-oriented documents* (e.g. PDF, Doc, etc.) or multimedia documents
 - recognizing and extracting ontology objects
 - automatically learning of wrappers and learning of ontology instances [Staab et Al.] by exploiting spatial patterns
 - navigating and accessing Deep Web data sources and dynamic components

Thank
you



For Further Reading I

-  S. Adali, M. L. Sapino, and V. S. Subrahmanian.
An algebra for creating and querying multimedia presentations.
Multimedia Syst., 8(3):212–230, 2000.
-  P. Balbiani, J.-F. Condotta, and L. F. d. Cerro.
A model for reasoning about bidimensional temporal relations.
In *Proc. of KR-2008*, pages 124–130, 1998.
-  R. Baumgartner, S. Flesca, and G. Gottlob.
Visual web information extraction with lixto.
In *VLDB*, pages 119–128, San Francisco, CA, USA, 2001.
Morgan Kaufmann Publishers Inc.

For Further Reading II

-  M. Benedikt and C. Koch.
Xpath leashed.
ACM Comput. Surv., 41(1):1–54, 2008.
-  G. Gottlob, C. Koch, and R. Pichler.
Efficient algorithms for processing xpath queries.
In *VLDB*, pages 95–106, 2002.
-  G. Gottlob, C. Koch, R. Pichler, and L. Segoufin.
The complexity of xpath query evaluation and xml typing.
J. ACM, 52(2):284–335, 2005.
-  J. Kong, K. Zhang, and X. Zeng.
Spatial graph grammars for graphical user interfaces.
ACM Trans. Comput.-Hum. Interact., 13(2):268–307, 2006

For Further Reading III



S. Mir, S. Staab, and I. Rojas.

Unsupervised approach for acquiring ontologies and rdf data from online life science databases.

In *ESWC*, 2010.



A. Sahuguet and F. Azavant.

Building intelligent web applications using lightweight wrappers.

DKE, 36(3):283–316, 2001.



P. Wadler.

Two semantics for xpath.

Draft: <http://homepages>

.inf.ed.ac.uk/~wadler/papers/xpath-semantics, 2000.



For Further Reading IV



Y. Zhai and B. Liu.

Extracting web data using instance-based learning.

In *WWW*, pages 113–132, 2007.

Example 2

Acquiring the table in the document as a set of triples of the form
 <row-header, column-header, value>.

Change									
Direct Energy Content [TJ]	1994	1996	1998	2000	2004	2005	2006	2007	'94 - '07
Total Gross Electricity Production	148 708	192 879	147 998	129 776	145 583	130 468	164 199	140 964	-2.4%
Oil	9 547	20 808	17 906	15 964	5 881	4 933	5 811	4 616	-51.7%
- Crude Oil	-	-	-	-	7	-	-	-	*
Natural Gas	8 206	20 442	29 260	31 189	35 897	31 606	33 903	24 886	20%
Coal	119 844	142 795	85 151	60 022	67 232	55 665	88 439	71 631	-40.2%
Surplus Heat	-	123	136	139	40	-	-	-	*
Waste, non-renewable	463	610	702	994	1 183	1 459	1 472	1 416	20%
Renewable Energy	6 647	8 101	14 844	21 658	35 459	36 805	34 574	38 415	479%
Solar Energy	0	1	1	4	7	8	8	9	301%
Wind Power	4 093	4 417	10 152	15 268	23 696	23 810	21 989	25 823	531%
Hydro Power	117	69	98	109	95	81	84	101	-14.3%
Biomass	2 116	3 207	3 911	4 936	10 646	11 889	11 517	11 504	444%
- Straw	293	748	960	654	3 057	3 088	3 339	3 185	908%
- Wood	429	340	512	828	3 546	3 730	3 041	3 398	691%
- Waste, renewable	1 393	2 120	2 439	3 454	4 043	5 071	5 117	4 921	253%
Biogas	321	407	662	751	1 013	1 017	976	978	20%

```
for $rh in document ("table.pdf")
(2.1) //text [not(W::*)]
  return
<table-triples>
  {
    for $ch at $j in document ("table.pdf")
(2.2) //text [not(N::*)]
```

```
<row-header>
(2.3) {$rh}
</row-header>
<column-header>
(2.4) {$ch}
</column-header>
<value>
(2.5) {$rh/E::text [W,$j]}
</value>
}
</table-triples>
```

Core XPath

Definition

The syntax of Core XPath is defined by the following EBNF grammar

```
locpath      ::= '/' locpath | locpath '/' locpath |  
                  locpath '!' locpath | locstep.  
locstep      ::= axis '::' t | locstep '[' bexpr ']'  
bexpr        ::= bexpr 'and' bexpr | bexpr 'or' bexpr |  
                  'not(' bexpr ')' | locpath.  
axis         ::= xpathAxis | spatialAxis.  
xpathAxis    ::= 'self' | 'child' | 'parent' |  
                  'descendant' | 'descendant-or-self' |  
                  'ancestor' | 'ancestor-or-self' |  
                  'following' | 'following-sibling' |  
                  'preceding' | 'preceding-sibling'.  
spatialAxis ::= topAxis | dirAxis.  
topAxis      ::= 'EQ' | 'CD' | 'CR'.  
dirAxis      ::= 'B' | ... | 'U'.
```

Spatial Wadler Fragment

Definition

The syntax of the SWF-Queries is defined by the Core SXPath grammar with the following extensions.

```
expr      ::= locpath | bexpr | nexpr
dirAxis   ::= 'B' | ... | 'U' | disjDirAxis.
bexpr     ::= bexpr 'and' bexpr | bexpr 'or' bexpr |
              'not(' bexpr ')' | nexpr relop nexpr | 
              sexpr relop sexpr | locpath |
              locpath relop sexpr |
              locpath relop number.
nexpr     ::= number | nexpr arithop nexpr.
              'position()' | 'last()' | 'posFromS()' | 'lastFromS()' |
              'posFromN()' | 'lastFromN()' | 'posFromW()' | 'lastFromW()' |
              'posFromE()' | 'lastFromE()' | 'posSpatialNesting()'.
sexpr     ::= string.
arithop   ::= '+' | '-' | '*' | 'div' | 'mod'.
relop     ::= '=' | '!= | '<' | '<=' | '>' | '>='.
```

Input: A set of nodes Γ and an axis $\chi \in \Delta$

Output: $\chi(\Gamma)$

Method: $\text{eval}_\chi(\Gamma)$

- (1.1) **function** $\text{eval}_{\text{self}}(\Gamma) := \Gamma.$
- (1.2) **function** $\text{eval}_{\chi_t}(\Gamma) := \text{eval}_{E(\chi_t)}(\Gamma).$
- (1.3) **function** $\text{eval}_{\chi_s}(\Gamma) := \text{eval}_{\{\rho_i \mid \rho_i \in \mu(\chi_s)\}}(\Gamma).$
- (1.4) **function** $\text{eval}_{\chi_s^{-1}}(\Gamma) := \text{eval}_{\{\rho_i^{-1} \mid \rho_i \in \mu(\chi_s)\}}(\Gamma).$
- (1.5) **function** $\text{eval}_\varrho(\Gamma)$ **begin**
- (1.6) $\Gamma' := \emptyset;$
- (1.7) **foreach** $u \in \Gamma \cap u \in V_V$ **do**
- (1.8) **foreach** $\rho_i \in \varrho$ **do**
- (1.9) $\Gamma' := \Gamma' \cup_{\text{set}} f_{\rho_i}(u)$ **od od**
- (1.10) **return** Γ' **end.**

(Location step evaluation algorithm)

Input: A set of nodes Γ and a location step $e = \chi :: \tau[e_1] \dots [e_m]$

Output: $P[e](\Gamma)$

Method: $eval(e, \Gamma)$ **begin**

(2.1) $Res := \emptyset$

(2.2) $W := \chi(\Gamma) \cap T(\tau);$

(2.3) **for** each $u \in \Gamma$ **do**

(2.4) $W' := \{w \mid w \in W \wedge u \chi w\}$

(2.5) **for** each e_i with $1 \leq i \leq m$ (in ascending order) **do**

(2.6) $\vec{W} := layering(W')$

(2.7) $W' := \{w \mid w \in \vec{W} \wedge \varepsilon[e_i](\vec{c}_w) = true \wedge$
 $\vec{c}_w := \langle w, idx_\chi(w, \vec{W}), |\vec{W}|, pidx_{\leq t}(w, \vec{W}), plast_{\leq t}(\vec{W}),$
 $pidx_{\leq \rightarrow}(w, \vec{W}), plast_{\leq \rightarrow}(\vec{W}), pidx_{\leq \leftarrow}(w, \vec{W}), plast_{\leq \leftarrow}(\vec{W}),$
 $pidx_{\leq \leftarrow \rightarrow}(w, \vec{W}), plast_{\leq \leftarrow \rightarrow}(\vec{W}), pidx_{\leq t}(w, \vec{W}) \rangle\}$

od

(2.8) $Res := Res \cup W'$

od

(2.9) **return** Res **end;**