Corso di Informatica di base per le discipline umanistiche - XPATH

Francesca Frontini

Pavia AA 2008-2009

Materiale

http://www.w3schools.com/xpath/default.asp

Cartella "xpath esercizi" scaricabile dalla pagina di questo modulo
http://lettere.unipv.it/diplinguistica/pagina.php?id=180

- alcuni file .xml con rispettive .dtd
- un file output.html
- un file mieixpath.txt

Definizione di XPATH

XPATH è un linguaggio sviluppato per trovare informazioni in un documento XML

XPATH serve per navigare tra gli elementi e gli attributi di un documento XML

"PATH" = "percorso" !!!

XPATH è un linguaggio SEMPLICE!!!!

OSSERVIAMO QUESTO DOCUMENTO XML

Example

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<bookstore>
<book>
  <title lang="eng">Harry Potter</title>
  <price>29.99</price>
</book>
<book>
  <title lang="eng">Learning XML</title>
  <price>39.95</price>
</book>
</bookstore>
```

Example

QUESTO

```
/bookstore/book[price>35.00]/title
```

è un XPATH valido per il documento che abbiamo osservato.

Secondo voi, dove ci porta questo “percorso”?

COME SI USA

- dove si scrive un comando xpath?
- come si applica il comando ad un certo file xml?
- dove escono i risultati?

I Browser (IExplorer, Mozilla Firefox, ...) sono in grado di interpretare la sintassi xpath.

... con un po' di sforzo!

COME SI USA

Uno dei modi possibili è costruire un file .html in cui si dice al browser:

1) dove trovare il file xml da interrogare
2) quale xpath applicare
3) come visualizzare i risultati

1 e 2 sono passaggi facili, e 3..... è già fatto e non ce ne dobbiamo preoccupare!

COME SI USA

Si deve aprire il file **output.html**
(preferibilmente utilizzando il browser mozilla Firefox)

PROMPT: Quale file vuoi interrogare?

es: hamlet.xml (il file da interrogare deve trovarsi nella stessa cartella di output.html)

COME SI USA

PROMPT: Quale path vuoi applicare?

es: /PLAY/ACT[1]/SCENE[1]/STAGEDIR

MANTENERE SEMPRE output.html nella STESSA CARTELLA del documento xml da caricare.

UN PO' DI ORDINE!

Nella cartella in cui abbiamo messo l'xml e output.html, troviamo anche un file chiamato **mieixpath.txt**

In questo file trovate gli xpath che proveremo, il file xml interrogato e una descrizione di cosa fanno.

Qui salverete allo stesso modo anche quelli che proverete voi!

Per provare una nuova ricerca, ricaricare output.html

PROVIAMO

A questo punto, possiamo provare a vedere come funziona!

Ma prima .... apriamo il documento **hamlet.xml** con il browser per vederne la struttura.

(come vedete il browser non da errori, quindi il documento è valido rispetto alla sua dtd; se volete potete osservare anche la dtd, aprendo play.dtd con blocco note!)

PROVIAMO

Secondo voi dove porta questo percorso?

Example

```
ESEMPIO 1

/PLAY/ACT[1]/SCENE[1]/STAGEDIR
```

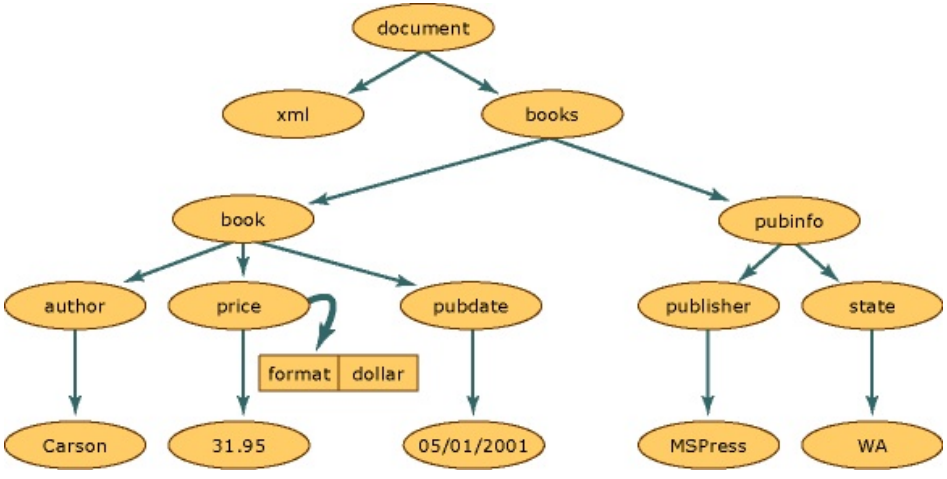
Ora aprite output.html e applicate il path

Cosa succede?

NODI

Ora che abbiamo un'idea di cos'è l'xpath e di come si usa, studiamo un po' la sintassi

COSA SONO I NODI



COME SI SELEZIONANO I NODI

“nome del nodo” seleziona tutti i nodi figli di quel nodo

/ seleziona a partire dal nodo radice

// MioNodo seleziona a partire da MioNodo

COME SI SELEZIONANO I NODI

Cosa succede se seleziono un nodo che ha al suo interno solo nodi figli, ma nessun testo non marcato?

Example

```
ESEMPIO 2

/PLAY
```

Non si vede niente! perché il browser mi mostra solo il contenuto del nodo, e questo nodo contiene solo altri nodi.

DOBBIAMO PERCORRERE L'ALBERO FINO ALLE FOGLIE PER AVERE UN OUTPUT A VIDEO!

COME SI SELEZIONANO I NODI

Example

```
ESEMPIO 3

/PLAY/ACT/SCENE/TITLE
```

oppure

Example

```
ESEMPIO 4

//SCENE/TITLE
```

OMONIMIA

Cosa succede quando ci sono due nodi con lo stesso nome a livelli diversi della gerarchia?

Example

```
ESEMPIO 5

//TITLE
```

cosa trova?

OPERATORI LOGICI

Per selezionare due alternative, posso usare l'operatore logico AND

Example

```
ESEMPIO 6

//ACT/TITLE | //SCENE/TITLE
```

SELEZIONARE UN SOLO NODO

A) contare

Example

```
ESEMPIO 7

/PLAY/ACT[3]/SCENE[1]/SPEECH[19]/LINE[1]
```

attenzione... IEXPLORER parte a contare da 0, altri browser da 1
quindi per EXPLORER bisogna diminuire il contatore di una unità

SELEZIONARE UN SOLO NODO

B) chiamare per nome

Example

```
ESEMPIO 8

/PLAY/ACT[3]/SCENE[1]/SPEECH[SPEAKER='HAMLET']/LINE
```

SELEZIONARE NODI SCONOSCIUTI

Possiamo selezionare tutti i figli di un certo nodo usando il carattere jolly *

Example

ESEMPIO 9

```
/PLAY/ACT/SCENE/SPEECH/*
```

SELEZIONARE NODI SCONOSCIUTI

Come fare selezionare tutta la tragedia?

Example

```
ESEMPIO 10

/PLAY/*"
```

Il comando seleziona i figli di PLAY e ne stampa il contenuto..

ma alcuni dei figli di PLAY non sono terminali, quindi ho un errore!

SELEZIONARE NODI SCONOSCIUTI

Posso aggiungere asterischi...

Example

```
ESEMPIO 11

/PLAY/*/*/*/*
```

... ma non vedrò mai tutto il testo della tragedia, perché non tutti i nodi terminali sono allo stesso livello. Che fare?

AXES

Gli AXES definiscono una relazione di parentela tra i nodi;

sono formati da un nome (relazione), un nodo test (a cosa applico la relazione) e un predicato (non obbligatorio)

Example

```
ESEMPIO 12


axisname::nodetest[predicate]
```

AXES

La relazione “descendant” identifica tutti i discendenti di un certo nodo

Example

```
ESEMPIO 13

/PLAY/descendant::*
```

ATTRIBUTI

I nodi possono avere anche degli **attributi**.

Come ci riferiamo all'attributo di un nodo?

Usiamo un testo annotato con attributi:

varney.xml contiene i primi 4 capitoli di un romanzo chiamato *Varney, the vampire*

ATTRIBUTI

Il testo contiene annotazione morfosintattica:

Parti del Discorso “pos”

Lemmi “hw”

sono espressi come attributi del nodo “w”

Aprite **varney.xml** con il browser e guardate come è costruito

(FACOLTATIVO: non c’è una dtd, ma se ne può sempre scrivere una ....!)

Come ci si riferisce ad un attributo?

Example

```
ESEMPIO 13

varney.xml

//w[@pos='JJ']
```

JJ = aggettivo; in questo modo posso estrarre tutti gli aggettivi del testo.

Come potete notare gli aggettivi ci dicono molto sul carattere di questo libro...

FACOLTATIVO.. copiare tutti gli aggettivi in un foglio di calcolo, e ordinarli per frequenza!

I verbi sono etichettati con 3 etichette diverse: VVN per i partcipi e VVZ per i verbi finiti

come fare per **catturarli tutti**?

Example

```
ESEMPIO 14

//w[@pos='VVZ'] | //w[@pos='VVN']"
```

E se voglio il **lemma**?

Example

```
ESEMPIO 15

//w[@pos='VVZ']/@hw"
```