XQuery Semantics

CSE 232B

March 31, 2016

1 The XPath Sub-language of XQuery

We consider XPath, the sublanguage of XQuery which deals with specifying paths along which the XML tree is to be navigated to extract data from it.[1]

Any expression generated by the following context-free grammar is a valid XPath expression.

extract text as children

$$
\begin{array}{rrcl}
\text{(absolute path)} & ap & \to & \mathsf{doc}(\mathit{fileName})/rp \\
 & & | & \mathsf{doc}(\mathit{fileName})//rp \\
 & & & \\
\text{(relative path)} & rp & \to & \mathit{tagName} \mid * \mid . \mid .. \mid \mathsf{text}() \mid @\text{attName} \\
 & & | & (rp) \mid rp_1/rp_2 \mid rp_1//rp_2 \mid rp[f] \mid rp_1, rp_2 \\
 & & & \\
\text{(path filter)} & f & \to & rp \mid rp_1 = rp_2 \mid rp_1 \ \mathsf{eq} \ rp_2 \mid rp_1 == rp_2 \mid rp_1 \ \mathsf{is} \ rp_2 \\
 & & | & (f) \mid f_1 \ \mathsf{and} \ f_2 \mid f_1 \ \mathsf{or} \ f_2 \mid \mathsf{not} \ f
\end{array}
$$

The above grammar only helps us check whether an XPath expression p has the correct syntax. But what is its **meaning**, i.e. what is the result of extracting from an XML tree the data reachable by navigating along p? To answer this question, we need to settle the following problem. How can one define the meaning of *any* XPath expression without explicitly listing each such expression and, for each possible XML document, the associated result? Note that this would be an unfeasible approach, as there are infinitely many XPath expressions, as well as infinitely many XML trees.

The solution is a standard one, adopted from programming language theory. We will define a function which, applied to any XPath expression p and XML tree rooted at node n, will return the list of nodes reachable by navigating along p. Recall that we consider two kinds of nodes in the XML tree: *element* nodes, and *text* nodes. Text nodes may be associated to element nodes.

We will use the following functions

[1] For the sake of simplicity, we will only consider a restriction of the full W3C XPath standard.

function	returns
$[\![ap]\!]_A$	the list of (element or text) nodes reached by navigating from the root along absolute path ap
$[\![rp]\!]_R(n)$	the list of (element or text) nodes reachable from element node n by navigating along the path specified by relative XPath expression rp.
$[\![f]\!]_F(n)$	true if and only if the filter f holds at node n
$\mathsf{root}(fn)$	the root of the XML tree corresponding to the document fn
$\mathsf{children}(n)$	the list of children of element node n, ordered according to the document order
$\mathsf{parent}(n)$	a singleton list containing the parent of element node n, if n has a parent. The empty list otherwise.
$\mathsf{tag}(n)$	the tag labeling element node n
$\mathsf{txt}(n)$	the text node associated to element node n

List manipulations We will also use the following notation on list manipulations. $< a, b, c >$ denotes a list of three entries (a is the first, c the last). $<>$ denotes the empty list, and $< e >$ is the singleton list with unique entry e.

In the following, l_1, l_2 are the lists $l_1 =< x_1, \ldots, x_n >$ and $l_2 =< y_1, \ldots, y_m >$.

$$l_1, l_2$$

denotes the concatenation of the two lists, i.e. the list $< x_1, \ldots, x_n, y_1, \ldots, y_m >$.

$$\mathsf{unique}(l_1)$$

denotes the list obtained by scanning l from head to tail and removing any duplicate elements that have been previously encountered.

For example, $< 1, 2, 3 >, < 2, 3, 4 >=< 1, 2, 3, 2, 3, 4 >$, and $\mathsf{unique}(< 1, 2, 3 >, < 2, 3, 4 >) =< 1, 2, 3, 4 >$.

The notation $< f(x) \mid x \leftarrow l_1 >$ is called a *list comprehension*, and it is shorthand for a loop which binds variable x in order against the entries of l_1, and returns the list with entries given by applying f to each binding of x:

$$< f(x) \mid x \leftarrow l_1 >=< f(x_1), \ldots, f(x_n) >$$

A list comprehension can have arbitrarily many condition and variable binding expressions. In general, if $c(v_1, \ldots, v_k)$ is a condition involving variables v_1 through v_k,

$$< f(v_1, \ldots, v_k) \mid v_1 \leftarrow l_1, \ldots, v_k \leftarrow l_k, c(v_1, v_2, \ldots, v_k) >$$

is short for the function defined by the following pseudocode fragment:

```
result := <>
foreach v1 in l1
 ...
   foreach vk in lk
     if c(v1,...,vk) then
       result := result, <f(v1,...,vk)>
return result
```

We are now ready to define the meaning of an XPath expression.

$$[\![\mathsf{doc}(\mathit{fileName})/rp]\!]_A = [\![rp]\!]_R(\mathsf{root}(\mathit{fileName})) \tag{1}$$
$$[\![\mathsf{doc}(\mathit{fileName})//rp]\!]_A = [\![.//rp]\!]_R(\mathsf{root}(\mathit{fileName})) \tag{2}$$

$$[\![tagName]\!]_R(n) = < c \mid x \leftarrow [\![*]\!]_R(n), \mathsf{tag}(n) = tagName > \tag{3}$$
$$[\![*]\!]_R(n) = \mathsf{children(n)} \tag{4}$$
$$[\![.]\!]_R(n) = < n > \tag{5}$$
$$[\![..]\!]_R(n) = \mathsf{parent}(n) \tag{6}$$
$$[\![\mathsf{text()}]\!]_R(n) = \mathsf{txt}(n) \tag{7}$$
$$[\![@attName]\!]_R(n) = \mathsf{attrib}(n, attName) \tag{8}$$
$$[\![(rp)]\!]_R(n) = [\![rp]\!]_R(n) \tag{9}$$
$$[\![rp_1/rp_2]\!]_R(n) = \mathsf{unique}(< y \mid x \leftarrow [\![rp_1]\!]_R(n), y \leftarrow [\![rp_2]\!]_R(x) >) \tag{10}$$
$$[\![rp_1//rp_2]\!]_R(n) = \mathsf{unique}([\![rp_1/rp_2]\!]_R(n), [\![rp_1/*//rp_2]\!]_R(n)) \tag{11}$$
$$[\![rp[f]]\!]_R(n) = < x \mid x \leftarrow [\![rp]\!]_R(n), [\![f]\!]_F(x) > \tag{12}$$
$$[\![rp_1, rp_2]\!]_R(n) = [\![rp_1]\!]_R(n), [\![rp_2]\!]_R(n) \tag{13}$$

$$[\![rp]\!]_F(n) = [\![rp]\!]_R(n) \neq <> \tag{14}$$
$$[\![rp_1 = rp_2]\!]_F(n) = [\![rp_1 \;\mathsf{eq}\; rp_2]\!]_F(n) = \exists x \in [\![rp_1]\!]_R(n)\ \exists y \in [\![rp_2]\!]_R(n)\ x \;\mathsf{eq}\; y \tag{15}$$
$$[\![rp_1 == rp_2]\!]_F(n) = [\![rp_1 \;\mathsf{is}\; rp_2]\!]_F(n) = \exists x \in [\![rp_1]\!]_R(n)\ \exists y \in [\![rp_2]\!]_R(n)\ x \;\mathsf{is}\; y \tag{16}$$
$$[\![(f)]\!]_F(n) = [\![f]\!]_F(n) \tag{17}$$
$$[\![f_1 \;\mathsf{and}\; f_2]\!]_F(n) = [\![f_1]\!]_F(n) \wedge [\![f_2]\!]_F(n) \tag{18}$$
$$[\![f_1 \;\mathsf{or}\; f_2]\!]_F(n) = [\![f_1]\!]_F(n) \vee [\![f_2]\!](n) \tag{19}$$
$$[\![\mathsf{not}\; f]\!]_F(n) = \neg [\![f]\!]_F(n) \tag{20}$$

Value-based and Identity-based Equality XPath distinguishes among two types of equality. Two XML nodes n and m are *value-equal* (denoted n eq m or $n = m$) if and only if the trees rooted at them are isomorphic. That is, if

- $\mathsf{tag}(n) = \mathsf{tag}(m)$ and
- $\mathsf{text}(n) = \mathsf{text}(m)$ and
- n has as many children as m and
- for each k, the k^{th} child of n and the k^{th} child of m are value-equal.

In other words, n is a copy of m. n and m are *id-equal* (denoted n is m or $n == m$) if and only if they are identical. That is, a node n is only id-equal to itself. n is not id-equal to a distinct copy of itself. Note that id-equality implies value-equality, but not viceversa.

2 The XQuery Sub-language for the Project

The W3C XQuery standard contains many bells and whistles which we will abstract from for the sake of simplicity. For our purposes, the syntax of XQuery is defined as follows:

$$
\begin{array}{llcl}
\text{(XQuery)} & XQ & \rightarrow & Var \mid StringConstant \mid ap \\
 & & \mid & (XQ_1) \mid XQ_1, XQ_2 \mid XQ_1/rp \mid XQ_1//rp \\
 & & \mid & \langle tagName \rangle \{XQ_1\} \langle /tagName \rangle \\
 & & \mid & forClause\ letClause\ whereClause\ returnClause \\
 & & \mid & letClause\ XQ_1 \\
 & & & \\
 & forClause & \rightarrow & \mathsf{for}\ Var_1\ \mathsf{in}\ XQ_1, Var_2\ \mathsf{in}\ XQ_2, \ldots, Var_n\ \mathsf{in}\ XQ_n \\
 & & & \\
 & letClause & \rightarrow & \epsilon \mid \mathsf{let}\ Var_{n+1} := XQ_{n+1}, \ldots, Var_{n+k} := XQ_{n+k} \\
 & & & \\
 & whereClause & \rightarrow & \epsilon \mid \mathsf{where}\ Cond \\
 & & & \\
 & returnClause & \rightarrow & \mathsf{return}\ XQ_1 \\
 & & & \\
 & Cond & \rightarrow & XQ_1 = XQ_2 \mid XQ_1\ \mathsf{eq}\ XQ_2 \\
 & & \mid & XQ_1 == XQ_2 \mid XQ_1\ \mathsf{is}\ XQ_2 \\
 & & \mid & \mathsf{empty}(XQ_1) \\
 & & \mid & \mathsf{some}\ Var_1\ \mathsf{in}\ XQ_1, \ldots, Var_n\ \mathsf{in}\ XQ_n\ \mathsf{satisfies}\ Cond \\
 & & \mid & (Cond_1) \mid Cond_1\ \mathsf{and}\ Cond_2 \mid Cond_1\ \mathsf{or}\ Cond_2 \mid \mathsf{not}\ Cond_1
\end{array}
$$

Element and Text Node Constructors We will use the function

$$\mathsf{makeElem}(t, l)$$

which takes as arguments a tag name t and a (potentially empty) list of XML nodes l and returns a new XML element node n with $\mathsf{tag}(n) = t$ and $\mathsf{children}(n)$ a list of *copies* of the nodes in l (these are *deep* copies, i.e. the entire subtrees rooted at these nodes are copied as well. . Similarly,

$$\mathsf{makeText}(s)$$

takes as argument a string constant s and returns an XML text node with value s.

Variable Scope As in any programming language with variables, we need to define the scope of variables. We first note that variables can be defined only by for, let and some clauses. We impose the following scoping rules, which are quite natural for any programming language with block structure.

- The scope of variables bound in a for clause extends to the corresponding (as given by production 8 of non-terminal XQ above) let clause (if any), where clause (if any) and return clause.
- The scope of the variables bound in a let clause extends to the following where and return clauses (if the applicable production is no. 8 above), or to the XQ_1 (if the applicable production is no. 9 above).
- The scope of the variables bound in a some clause extends to the condition in the satisfies clause.
- Moreover, within any for, let or some clause, every XQ_i used to bind variable Var_i may depend on the previously defined variables.

A definition of variable v will override within the definition's scope any prior definition of variable v. For instance, in a query

$$\mathsf{for}\ v\ \mathsf{in}\ XQ_1,\ w\ \mathsf{in}\ XQ_2\ \mathsf{let}\ v := XQ_3\ \mathsf{where}\ Cond\ \mathsf{return}\ XQ_4$$

any reference to v in *Cond* and XQ_4 refers to the definition using XQ_3, while any reference in XQ_2 refers to the definition using XQ_1.

Evaluating Expressions with Free Variables in a Context Since we intend to evaluate an expression by evaluating its sub-expressions first, we need to cover the case when the sub-expression mentions free variables defined outside. To this end, we will record all variable bindings in an auxiliary data structure called a *context*, and pass the context as argument to the evaluation function, which will look up prior variable bindings in the context. Think of a contet as an associative array which relates variables to the value they are bound to. A context supports two operations:

- $\{Var \mapsto v\}C$ extends the context C with a new binding for variable *Var* to value v. This operation has no side-effect, i.e. it does not change C, instead returning a brand new context which copies from C all bindings of variables other than *Var*.
- $C(Var)$ is the operation of looking up the binding of variable Var in context C, yielding the value *Var* was bound to.[2]

To support the override rule for variable definitions, we require any context to behave as follows:

$$(\{Var \mapsto u\}C)(Var) = u$$

which implies in particular (for $C = \{Var \mapsto v\}C'$) that

$$(\{Var \mapsto u\}\{Var \mapsto v\}C')(Var) = u.$$

The Evaluation Functions The function evaluating an XQuery expression XQ within a context C is $[\![XQ]\!]_X(C)$, and it returns a list of element and text nodes. The function evaluating a condition *Cond* within a context C is $[\![Cond]\!]_C(C)$ and it evaluates to a boolean. We define the two functions below.

$$
\begin{array}{rcll}
[\![Var]\!]_X(C) & = & <\ C(Var)\ > & (21)\\
[\![StringConstant]\!]_X(C) & = & <\mathsf{makeText}(StringConstant)> & (22)\\
[\![ap]\!]_X(C) & = & [\![ap]\!]_A & (23)\\
[\![(XQ_1)]\!]_X(C) & = & [\![XQ_1]\!]_X(C) & (24)\\
[\![XQ_1, XQ_2]\!]_X(C) & = & [\![XQ_1]\!]_X(C), [\![XQ_2]\!]_X(C) & (25)\\
[\![XQ_1/rp]\!]_X(C) & = & \mathsf{unique}(< m \mid n \leftarrow [\![XQ_1]\!]_X(C), m \leftarrow [\![rp]\!]_R(n) >) & (26)\\
[\![XQ_1//rp]\!]_X(C) & = & \mathsf{unique}(< m \mid n \leftarrow [\![XQ_1]\!]_X(C), m \leftarrow [\![.//rp]\!]_R(n) >) & (27)\\
[\![\langle tagName\rangle\{XQ_1\}\langle/tagName\rangle]\!]_X(C) & = & <\mathsf{makeElem}(tagName, [\![XQ_1]\!]_X(C))> & (28)\\
\\
[\![XQ_1\ \mathsf{eq}\ XQ_2]\!]_C(C) = [\![XQ_1 = XQ_2]\!]_C(C) & = & \exists x \in [\![XQ_1]\!]_X(C)\ \exists y \in [\![XQ_2]\!]_X(C)\ x\ \mathsf{eq}\ y & (29)\\
[\![XQ_1\ \mathsf{is}\ XQ_2]\!]_C(C) = [\![XQ_1 == XQ_2]\!]_C(C) & = & \exists x \in [\![XQ_1]\!]_X(C)\ \exists y \in [\![XQ_2]\!]_X(C)\ x\ \mathsf{is}\ y & (30)\\
[\![\mathsf{empty}(XQ_1)]\!]_C(C) & = & [\![XQ_1]\!]_X(C) = <> & (31)\\
\left[\!\!\left[\begin{array}{l}\mathsf{some}\ Var_1\ \mathsf{in}\ XQ_1, \ldots, Var_n\ \mathsf{in}\ XQ_n\\ \mathsf{satisfies}\ Cond\end{array}\right]\!\!\right]_{Cn}(C) & = & \exists v_1 \in [\![XQ_1]\!]_X(C_0) & \\
& & \ldots & \\
& & \exists v_n \in [\![XQ_n]\!]_X(C_{n-1}) & \\
& & [\![Cond]\!]_C(C_n) & (32)
\end{array}
$$

where $C_0 := C,\ C_i := \{Var_i \mapsto v_i\}C_{i-1},\ i \in [1, \ldots, n]$

[2]We shall assume that variables are always defined before being used (this can be easily checked at parsing time) and therefore define the evaluation only for well-formed XQuery expressions.

$$
\begin{aligned}
[\![(Cond_1)]\!]_C(C) &= [\![Cond_1]\!]_C(C) && (33)\\
[\![Cond_1 \ \mathsf{and}\ Cond_2]\!]_C(C) &= [\![Cond_1]\!]_C(C) \wedge [\![Cond_2]\!]_C(C) && (34)\\
[\![Cond_1 \ \mathsf{or}\ Cond_2]\!]_C(C) &= [\![Cond_1]\!]_C(C) \vee [\![Cond_2]\!]_C(C) && (35)\\
[\![\mathsf{not}\ Cond_1]\!]_C(C) &= \neg[\![Cond_1]\!]_C(C) && (36)
\end{aligned}
$$

Finally, we have

$$
\left[\!\!\left[\begin{array}{ll} \mathsf{let} & Var_1 \ := \ XQ_1, \ldots, Var_n \ := \ XQ_n \\ XQ_{n+1} & \end{array} \right]\!\!\right]_X (C) \quad = \quad [\![XQ_{n+1}]\!]_X(C_n) \tag{37}
$$

where $C_0 := C,\ C_i := \{Var_i \mapsto [\![XQ_i]\!]_X(C_{i-1})\}C_{i-1},\ i \in [1, \ldots, n]$

(38)

$$
\left[\!\!\left[\begin{array}{ll} \mathsf{for} & Var_1 \ \mathsf{in}\ XQ_1, \ldots, \\ & Var_n \ \mathsf{in}\ XQ_n \\ \mathsf{let} & Var_{n+1} := XQ_{n+1}, \ldots, \\ & Var_{n+k} := XQ_{n+k} \\ \mathsf{where} & Cond \\ \mathsf{return} & XQ_{n+k+1} \end{array} \right]\!\!\right]_X (C) \quad = \quad \begin{array}{ll} < & [\![XQ_{n+k+1}]\!]_X(C_{n+k}) \mid \\ & v_1 \leftarrow [\![XQ_1]\!]_X(C_0), \\ & \ldots, \\ & v_n \leftarrow [\![XQ_n]\!]_X(C_{n-1}), \\ & [\![Cond]\!]_C(C_{n+k}) \ > \end{array} \tag{39}
$$

where $C_0 := C,\ C_i := \{Var_i \mapsto v_i\}C_{i-1},\ i \in [1, \ldots, n]$

and $C_j := \{Var_j \mapsto [\![XQ_j]\!]_X(C_{j-1})\}C_{j-1},\ j \in [n+1, \ldots, n+k]$

Notice that the effect of the let construct is simply that of extending the context with bindings for the variables declred in the construct.