A Machine Learning Approach to Rapid Development of XML Mapping Queries

Atsuyuki Morishima
Hiroyuki Kitagawa
Akira Matsumoto
University of Tsukuba
mori@slis.tsukuba.ac.jp kitagawa@is.tsukuba.ac.jp akey@kde.is.tsukuba.ac.jp

Abstract

This paper presents XLearner, a novel tool that helps the rapid development of XML mapping queries written in XQuery. XLearner is novel in that it learns XQuery queries consistent with given examples (fragments) of intended query results. XLearner combines known learning techniques, incorporates mechanisms to cope with issues specific to the XQuery learning context, and provides a systematic way for the semi-automatic development of queries. This paper describes the XLearner system. It presents algorithms for learning various classes of XQuery, shows that a minor extension gives the system a practical expressive power, and reports experimental results to demonstrate how XLearner outputs reasonably complicated queries with only a small number of interactions with the user.

1 Introduction

As the amount of XML data grows, the need to integrate and restructure XML data increases significantly. A major approach is to write XML queries to map one XML data structure into another, but developing such mapping queries requires tremendous effort. The effort is inevitable because of XML data characteristics and the nature of integration/restructuring: (1) XML data is nested, unnormalized, and an instance of semistructured data. This requires complex navigation and pattern matching of XML data. (2) Mapping XML data to a given complex schema requires complex construction of XML query results.

Since the development requires tremendous effort, tools to assist query development are beneficial. A well-known example is the theoretical tools for type checking [4], which checks if every output from an XML query is consistent with a given DTD. Another example is rapid development tools for XML mapping queries [14]. The latter tools are intended to play a similar role to rapid application development tools for programming languages; i.e., they are designed to help developers make their development processes more efficient. In general, such RAD tools provide various ways for the semi-automatic generation of code.

Our proposed tool, called XLearner, is a rapid query development tool. A key feature is that it uses machine learning techniques to generate complicated XQuery queries after a small number of simple interactions with the user. XLearner requires the user merely to (1) show the system examples (fragments) of the intended query results and (2) answer simple Yes or No questions posed by the system. In other words, XLearner learns XQuery queries if the user gives example fragments and answers to the simple questions posed by XLearner. It is worth noting that the questions do not involve query language jargon such as joins.

Development of XLearner is challenging for three reasons. First, learning XML queries through examples is anything but trivial: XML is an instance of semistructured data, whose data structure is often irregular and implicit. This is a striking difference from relational databases, where the data structure is completely regular and explicit. The structure of XML query results is deeply nested and complex in general, in contrast to relational databases where the result is always a flat table. XLearner combines known machine learning techniques, incorporates mechanisms to cope with issues specific to the XQuery context, and provides a non-adhoc framework to achieve such non-trivial tasks. We believe this is a significant contribution.

Second, naive approaches have problems with computational complexity. Many XML query languages support path regular expressions or their variations, such as XPath [19]. The expressive power of regular expressions is equivalent to that of finite automata1, and Gold has shown that the problem of finding a dfa of a minimum number of states compatible with a given set of examples is NP-complete [8]. It is known, however, that active learning [3] gives a polynomial-time algorithm to solve the problem, where active learning is a framework that allows the system (learner) to ask the user (teacher) questions. XLearner applies an active learning algorithm to our query learning context.

Finally, even with a polynomial-time algorithm, a tool is not necessarily practical. XLearner is based on an active learning framework, and requires interactions with the user. Keeping the number of interactions between XLearner and the user small is a critical point. If XLearner required several hundred interactions to learn a simple query, it would

1The expressive power of XPath is not equal to, but overlaps that of regular expressions.
not be a practical system. XLearner reduces the number of interactions in the following way: (1) It uses constraints specific to XML queries. (2) It requires the input of explicit query specifications in part. Specifically, the user is required to specify selection predicates on values. Because users usually know what their predicates are and teaching such predicates by example is more troublesome than specifying them, this provides a balanced combination for the efficient development of queries.

Development of XLearner also presents an interesting problem: “which class of XQuery query is learnable through examples?” We give algorithms for MAT-learning various classes of XQuery queries, which means that queries in the classes are learnable if the user is a minimally adequate teacher [2]. A minimally adequate teacher answers particular types of questions. In addition, we propose a minor extension that gives the system a practical expressive power to learn queries, including 19 of 20 queries in XMark [16] benchmark and 11 of 12 queries in XML Use Case [18] “XMP” (Experiences and Exemplars).

In summary, the contribution of this paper is as follows: (1) We propose XLearner, an intelligent tool for learning XQuery queries through given examples of query results. (2) We explain algorithms for learning various classes of XQuery queries in the active learning approach. (3) We show that a minor extension of the algorithm has practical expressive power. (4) We give experimental results to demonstrate that XLearner learns XQuery queries with a small number of interactions with the user.

Related Work. LSD [7] is a system that uses and extends machine learning techniques to semi-automatically create semantic mappings between a mediated schema and local schemas of data sources. It takes as input manual mappings between the mediated schema and a small set of data sources, and uses those mappings and data instances to propose mappings for subsequent data sources. With that capability, LSD is useful in a scenario where a data integrator plays a crucial role, who knows (1) how other data sources are mapped to the mediated schema and (2) what data is stored in the already-mapped data sources. On the other hand, XLearner requires no such information and uses other kinds of inputs. We believe XLearner is useful in another common scenario where each data provider plays a crucial role, who has to map their data to a given common schema.

Clio [14] is a well-known tool for the semi-automatic generation of schema-mapping queries written in SQL and XQuery. While XLearner and Clio look similar, they have different assumptions and different mechanisms to achieve the same goal. We believe future query development tools benefit from both approaches. Clio’s code generation is based on value-correspondences given by the user and semantic constraints given by schemas. Clio generates query candidates by taking advantage of the assumption that the result of the generated query retains semantic constraints given by the source schema. XLearner uses machine learning techniques to learn queries from interactions with the user, and there is no default assumption on how to generate XML instances for the given target schema. Moreover, Clio assumes that the source XML has a regular structure, in the sense that generated queries contain only simple location paths (such as /a/b/c). This is a significant difference; XLearner can generate xpath expressions corresponding to regular languages, which contributes to the ability to learn a large class of queries.

There are theoretical studies in the field of computational learning theory. Angluin [2] introduced the concept of active learning and showed that a dfa having the minimum number of states consistent with given examples is learnable in polynomial time. Inspired by Angluin’s work, researchers have been developing algorithms to learn a variety of patterns from data instances. Their focus is on learnability in polynomial time. Target patterns include different classes of formal languages [15] [9] and subclasses of CNF and DNF [6]. Some works are on extracting patterns from trees or semistructured data [1] [10] [5]. In general, studies in this area are purely theoretical and cover only simplified problems; the development of XLearner is not a variation of them. The reason is two-fold: First, learning XQuery queries implies learning compositions of various constructs of queries, including path expressions, join/selection conditions, and nested queries. Second, as mentioned, polynomial-time learning is not a sufficient condition for interactive systems. As far as we know, XLearner is the first system designed for learning practical XML queries from given examples.

2 XLearner

For an example scenario, we use XML data taken from XMark [16] whose structure is modeled after a database deployed by an Internet auction site. Figure 1(a) shows a fragment of its DTD. An item is an object that is for sale or has already been sold. Each item has a unique identifier and properties like name and description, all encoded as elements. Each item is assigned a world region represented by the item’s parent. A category has a unique identifier and a name. The categories are used to implement the classification of items, with incategory elements having IDREFs to the category elements. A closed auction represents an auction that is finished. Its properties include a reference to the item sold and the price. Figure 4(a) shows a fragment of the data instance.

Consider the XQuery query $q_1$ shown in Figure 2 (ignore the superscripts). The query outputs an XML instance that conforms to the schema shown in Figure 1(b), where each category element contains a set of items whose world regions are africa or europe and that were sold for less than 300 dollars.
There are two types of learner’s queries: membership queries and equivalence queries.

Membership Queries. To make a membership query for example $e$, XLearner chooses any XML node $n$ and asks the user whether $n$ is included in $EXT_e$. The user answers Yes (Y) or No (N). Figure 5 (a) shows an example of a membership query, which asks the user if the “XML book” under the $\texttt{asia}$ element is included in $EXT_{\text{H.Potter}}$. The answer
Types of queries is called a negative counterexample, because the sales price was more than $1000.

A teacher with the ability to give answers to the two types of queries is called a positive (negative) example. Dropped examples are positive examples by definition.

Equivalence Queries. The queries use hypotheses, i.e., intermediate results in the learning process. We use the hat notation \( \hat{\cdot} \) to differentiate a hypothesis from the correct answer \( x \) to learn. To make an equivalence query, XLearner highlights XML nodes in a hypothesis extent \( EXT_e \) in the XML browser and asks if \( EXT_e = EXT_t \). If true, the user clicks on the [OK] button. If not, the user gives a counterexample \( ce \) to XLearner. A counterexample is an XML node in the symmetric difference between \( EXT_e \) and \( EXT_t \). \( ce \) is called a positive (negative) counterexample. A positive (negative) counterexample is also a positive (negative) example by definition. Figure 5(b) shows an example of an equivalence query, where \( EXT_{\hat{\text{Potter}}} \) is highlighted in the browser. The user says “Encyclopedia” (with \( \text{id} \) “6”) is a negative counterexample, because the sales price was more than 300 dollars.

A teacher with the ability to give answers to the two types of queries is called a minimally adequate teacher [2].

Occasionally XLearner presents a dialog box (called a Condition Box) and requires explicit conditions. In the scenario, XLearner gives a Condition Box in the process of identifying \( EXT_{\hat{\text{Potter}}} \). The user drags and drops H. Potter’s price value (that is under closed_auction element) into the Condition Box and enters the selection condition \( <300 \) (Figure 5(c)).

If unable to find appropriate examples in a context, the user can change the context by switching to other choices of dropped examples to specify the same query. Details are omitted here to save space.

3 XQ-Tree

To discuss the learning algorithm, we introduce the XQ-Tree, our representation of XQuery queries. Figure 6 shows an XQ-Tree \( t_1 \) corresponding to \( q_1 \) in Figure 2. \( t_1 \) is essentially the same as \( q_1 \) except that (1) the nesting structure of \( flwr \) expressions is represented as a tree, and (2) it transforms every expression that returns a sequence of nodes (e.g. \( \text{Expr} \)) into the equivalent \( flwr \) expression (for \( \text{Expr} \)). Each node \( n \) in an XQ-Tree has the form \( \text{Expr} \) where \( \text{Expr} \) is a node identifier (We use Dewey encoding [13] here) and \( \text{Expr} \) is a query fragment (in the form of \( \text{flwr} \) expression, \( [\text{for} \text{Expr} \text{where} \text{Expr} \text{and} \text{Expr}] \) return \( \text{Expr} \), where \( [\ldots] \) is optional). Tags and node identifiers in \( \text{Expr} \) are omitted unless they are important.

The XQ-Tree representation has two important properties: (1) Every expression that returns a sequence of XML nodes (i.e., \( flwr \) expression) has a variable over the nodes in the sequence. (2) Edges between nodes suggest query dependencies among the nodes. For example, evaluation of \( q(n) \) depends on that of query fragments of \( n \)’s ancestors \( (\text{ancestors}_n) \) in general. We write \( \text{depends}_n \) to denote the set of all the nodes in \( t \) that \( n \) may depend on. As explained later, \( \text{depends}_n \) varies according to classes of queries. For the class to which the query in Figure 6 belongs, \( \text{depends}_n = \text{ancestors}_n \). If \( \text{depends}_n = \emptyset \) for some class, there is no possibility that \( q(n) \) has references to variables defined in other nodes.

For any XQ-Tree node \( n \) in \( t \), \( cq_n(n) \) denotes the complete query of \( n \). \( cq_n(n) \) is complete in the sense that it gives the complete computation of the XML node sequence corresponding to \( n \). For example, \( cq_n(N1.1.1) \) returns elements for all category names. Before defining \( cq_n(n) \), we define \( \text{compose}(q, q') \) as follows: Given query fragments \( q \) and \( q' \), the for (where) clause of \( \text{compose}(q, q') \) is the concatenation of \( \text{for} \) (where) clauses of \( q \) and \( q' \). Then, \( cq_n(n) \) is obtained by recursively applying function \( \text{compose}(q, q') \) to the query fragments in \( \{q(n)\} \cup \{q(m)\} m \in \text{depends}_n \). For example, \( cq_n(N1.1.1) = \text{for} \text{Expr} \text{where} \text{Expr} \text{return} \text{Expr} \) in the following, given a query fragment \( q(n) \).
Given a complete query \( q \), we write \( \{q\} \) to denote a set of XML nodes in \( q \)'s evaluation result. \( \{\{c_q_i (N.1.1)\}\} \) is a set of XML elements for category names. The subscript \( t \) is omitted when there is no ambiguity.

### 4 General Framework

Taking into account that XQuery queries are complicated with various query constructs, one of our technical contributions is that we have developed a general framework that includes several concepts (XQ-Trees, contexts, extents) to make possible the nontrivial task of learning XQuery queries from examples. In this section, we explain the proposed framework.

The learning process has two phases. First, XLearner creates an XQ-Tree skeleton \( t_s \) from a template and dropped examples. \( t_s \) gives the tree structure of the XQ-Tree to learn. Then, it traverses \( t_s \) and learns query fragment \( q(n) \) for each node \( n \) in \( t_s \). Intuitively, each \( q(n) \) in \( t_s \) is associated with one Drop Box in the template, and what XLearner learns through membership and equivalence queries on a Data Box is a path expression and selection conditions for the associated \( q(n) \).

#### 4.1 Template and XQ-Tree skeleton

As explained in Section 2, XLearner creates a template based on a given target schema. The template works as a basis for an XQ-Tree skeleton. Figure 7 (a) is the template for our example. A template is a tree, in which a node is created for each element type in the target schema. Each node has an element type with it. For example, an element definition \( A = (B, C, (D)E) \) results in a node \( n_A \) with four children \( n_B, n_C, n_D, n_E \). If an element has recursive definitions, a new node is created for each new occurrence of the same element type. Because it is possible that the template’s size is infinite, XLearner incrementally presents the template to the user. At first, only the first instance of recursively defined element is shown. When the user clicks on a node, it instantiates its children.

If the target schema guarantees that a node and its parent has one-to-one relationship, the edge connecting them is labeled as “1.” The labels are used in the learning process. For simplicity, the following discussions assume that each node has at most one child connected by 1-labeled edge and there are no consecutive 1-labeled edges on any path from the root to a leaf. (The template in Figure 7 satisfies this restriction.) It is easy to remove the restriction.

An XQ-Tree skeleton is a simple XQ-Tree computed from the template and the dropped example nodes. Figure 7 (b) illustrates the XQ-Tree skeleton for the example in Section 2. The tree structure is the minimum subtree of the template including all the nodes into which examples were dropped. Note that while the the size of the template can be infinite (if the target schema has recursive element definitions), every XQ-Tree skeleton is a tree with a finite number of nodes. For each XQ-Tree node \( n \), \( q(n) = \text{"return expr"} \). If an example node \( e \) is dropped into XQ-Tree node \( n, expr \), contains \( v_e \), which is a variable that corresponds to \( e \). (In Figure 7(b), \( v_{\text{book}} = cn, v_{\text{BestSeller}} = \text{in}, \) and \( v_{\text{BestSeller}} = d \).)

#### 4.2 Learning Each Query Fragment

We write \( n_e \) to denote an XQ-Tree node in the XQ-Tree skeleton \( t_s \), where \( e \) is the example node dropped into \( n_e \). XLearner learns the whole XQ-Tree by learning \( q(n_e) \) for each \( n_e \) in turn. XLearner assumes that \( q(n_e) \) has the form “for \( expr \) where \( expr_w \) return \( v_e \).” It is \( expr_w \) and \( expr_w \) that XLearner needs to learn for \( q(n_e) \). XLearner knows in advance the class of queries to learn so that it can compute \( \text{depends}(n_e) \) using \( t_s \). We give formal definitions of the context and extent first.

**Context.** We define the context of example \( e \) as follows: Let an assignment be a set of pairs \( \{(v,o)\} \) where \( v \) is a variable and \( o \) is an XML node. Then, given an XQ-Tree skeleton \( t_s \), a dropped example node \( e \), and the other dropped nodes for \( t_s \), we write \( \text{context}_t (e) \) to be an assignment:

\[
\{(v_e, e_i) | v_e, e_i \in n_e, e_i \in \text{depends}_t(n_e)\}
\]

where \( e_i \) is a dropped example node for XQ-Tree node \( n_e \), and \( v_{e_i} \) is a variable that corresponds to \( e_i \) in \( n_e \)’s return
5 Class X0 and its family

Given the general framework, challenging issues include how to traverse the XQ-Tree skeleton and how to combine known learning techniques for learning each query fragment. For explanation, we give an algorithm for a class of simple XQuery queries first and develop it for more general classes. This section gives algorithms for classes of XQuery queries without join/selection conditions.

X0, X0∗. X0∗ are classes of XQ-Trees where depends(n) = ϕ for every XQ-Tree node n. No where condition is allowed in queries in the classes. [X0] This is the simplest class of XQ-Trees with only one XQ-Tree node. An example XQ-Tree (and the corresponding query) in X0 is as follows:

Query: for $i$ in /site/regions/item return <result>$i</>

Before giving the X0’s definition, we define 0-Learnable(n), a predicate about learnability of a query fragment. The predicate holds when q(n) is learnable by a simple algorithm based on the Angluin’s algorithm [2]. Formally, 0-Learnable(n) holds iff q(n) = “for v in p return v” where p is a regular path expression that starts with a document function (i.e., the root of an XML document). Note that 0-Learnable(n) guarantees that depends(n) = ϕ and EXTc = {q(n)}.

Definition An XQ-Tree t is in Class X0 iff t consists of the root node n only and 0-Learnable(n).

Algorithm LEARN-X0. In X0, expr f = “v in p.” Therefore, EXT c ={[for v in p return v]}, and p is a path regular expression, which is the only thing to learn to compute EXT c. Because learning regular expressions is equivalent to learning deterministic finite automata, it is natural to use a learning technique to learn a dfa.

We apply the Angluin’s algorithm [2] to our context for efficiently learning a dfa from membership queries and equivalence queries. The algorithm first takes an example string, makes a number of membership queries on different strings, and constructs a hypothesis automaton $M$ consistent with the strings. It makes an equivalence query on $M$ to receive a counterexample (string) to modify $M$, and repeats the process until it finds the intended automaton. The key idea is to attempt to continually discover new states. Since a given counterexample suggests that it leads to a wrong state in the current hypothesis automaton, the algorithm discovers a new state. Figure 8 shows a step of the learning process where a given positive counterexample (that should be accepted) is used to find a new state for a dfa.

In our context, a sequence of tags, which represents a string, corresponds to a string in the Angluin’s algorithm. Figure 8 gives...
hypothesis automaton generated in a process of learning a DFA that corresponds to \( /site/region/asia \). The system first receives \( path(e) \) for the dropped example \( e \), where \( path(e) \) is the sequence of tags that matches the sequence of XML nodes from the XML instance’s root to \( e \). It then makes membership queries and equivalence queries so that it can learn \( \rho \). Note that the algorithm makes membership queries and equivalence queries on a sequence of tags, not an XML node. So XLearner has to choose an XML element \( \hat{m} \) s.t. \( path(m) = s \) to issue an membership query. To make an equivalence query, XLearner highlights XML nodes in \( EXT_e \).

[**X0**] This includes XQ-Trees with more than one node.

**Definition:** An XQ-Tree \( t \) is in **X0** iff for every node \( n \) in \( t \), \( 0\)-Learnable(\( n \)) holds.

Following is an example:

```xml
for $i$ in /site/regions/item/
return <result>$i</>
</>
N1:- for $i$ in /site/regions/item/
N1.1:- return $i{[N1.1]} /* result */
N1.1:- for $c$ in /site/categories/category/name
return $c{[N1.1]} /* cname */</>
</>
```

Every \( q(n) \) has a variable in its return clause. Therefore, every node in an XQ-Tree skeleton requires a dropped example in the learning process. XQ-Trees in the class correspond to nested queries whose query results are essentially Cartesian products.

**LEARN-X0**. **X0**’s definition says that for every node \( n_c \), \( depends(n_c) = \phi \) and \( EXT_e = \{|q(n_c)\}| \). This implies that XLearner can traverse the XQ-Tree skeleton in an arbitrary order and learn each query fragment using **LEARN-X0**.

[**X0**] In the previous classes, every \( q(n) \) is assumed to have a variable \( v_n \) in its return clause, because \( q(n) \) is learned using the example \( e \). We get **X0** when we relax the assumption and allow query fragments without variables under a certain condition. Following is an example:

```xml
for $c$ in /site/categories
return <root>
  for $c$ in /site/categories/category/name
  return $c{[N1.1]} /* cname */
</>
```

where “[1]” after a node identifier means the edge is labeled as 1 according to a given DTD. \( q(N1) \) and \( q(N1.1.1) \) have no variable in their return clause, but they are (indirectly) learnable because \( N1 \) and \( N1.1.1 \) satisfy a certain condition. Informally, such a node has to satisfy the following conditions (Figure 9): (1) If \( n \) has a child connected by 1-labeled edge (represented as \( C^1(n) \)), \( q(n) \) is learnable by applying **LEARN-X0** to \( collapse(n,C^1(n)) \). Here, \( collapse(n,n') \) be an XQ-Tree node whose query fragment, \( q(collapse(n,n')) \), is defined by \( compose(q(n),q(n')) \). This applies to \( N1 \), because \( 0\)-Learnable\((\text{collapse}(N1,N1.1.1)) \) holds. (2) Otherwise \( q(n) \) is just a holder of child XQ-Tree nodes. This applies to \( N1.1.1 \).

Formally, we say \( 0\)-Learnable(\( n \)) holds when both of the following conditions hold:

(A1) \( C^1(n) \) exists \( \Rightarrow 0\)-Learnable\((\text{collapse}(n,C^1(n))) \)

(A2) \( C^1(n) \) does not exist \( \Rightarrow q(n) = \text{return node id's of n's children} \).

**Definition:** An XQ-Tree \( t \) is in \( \textbf{X0}^+ \) iff for every node \( n \) in \( t \), \( 0\)-Learnable(\( n \)) \( \lor \) \( 0\)-Learnable(\( n \)).

**LEARN-X0**. This is the same as **LEARN-X0** except that it collapses nodes or returns an empty return clause when \( q(n) \) in the XQ-Tree skeleton has no variable in the return clause. Figure 9 shows how to learn each query fragment when an XQ-Tree skeleton is given to learn the query example above.

The result of **LEARN-X0** applied to the query example would be as follows:

```xml
for $c$ in /site/categories
return $c{[N1.1.1]} /* cname */
N1.1.1:- return $n{[N1.1.1.1]} /* name */
N1.1.1:- for $n$ in /site/name return $n
```

where \( N1 \) and \( N1.1 \) are collapsed into \( N1' \). Note that XQuery’s semantics guarantees that collapsing the nodes connected by 1-labeled edges does not change the query result.
6 Class X1 and its family

Class X1 and its family deal with selection conditions in where clauses. We introduce the concept of 1-learnability instead of 0-learnability and use it to define the classes.

Notations. When an XQ-Tree node has "for v in p," we write $expr(v)$ to denote the binding expression "v in p," and write $expr(v).path$ to denote the path expression p (e.g., $expr, \$(cn).path=\$c/name$ in Figure 6). $expr^*(v)$ is the closure of binding expressions to compute v. ($expr^*(_cn) = \{"\$c in /site/categories/category/\$c in \$/site/categories/category/name\} \}$ $expr^*(v)$ stands for a binding expression "v in p′," where p′ is the concatenation of all path expressions appearing in $expr^*(v)$. ($expr^*(_cn) = \{"\$c in \$/site/categories/category/name\} \}$ $expr^*(v)$ stands for a binding expression "v in p′," where p′ is the concatenation of all path expressions appearing in $expr^*(v)$. ($expr^*(_cn) = \{"\$c in \$/site/categories/category/name\} \}$. Let $n_v$ be the XQ-Tree node whose for clause defines v. Then, $associateable(v)$ is the set of variables appearing in $expr^*(v)$. ($associateable(_cn) = \{\$cn, \$ct\} \}$. Let $n_v$ be the XQ-Tree node whose for clause defines v. Then, $associateable(v)$ is the set of variables appearing in $expr^*(v)$. ($associateable(_cn) = \{\$cn, \$ct\}$. Let $n_v$ be the XQ-Tree node whose for clause defines v. Then, $associateable(v)$ is the set of variables appearing in $expr^*(v)$. ($associateable(_cn) = \{\$cn, \$ct\}$. Let $n_v$ be the XQ-Tree node whose for clause defines v. Then, $associateable(v)$ is the set of variables appearing in $expr^*(v)$. ($associateable(_cn) = \{\$cn, \$ct\}$. Let $n_v$ be the XQ-Tree node whose for clause defines v. Then, $associateable(v)$ is the set of variables appearing in $expr^*(v)$. ($associateable(_cn) = \{\$cn, \$ct\}$). The XQuery’s variable scope guarantees that (1) $associateable(v) \subseteq associateable(v)$, and (2) if $v \notin associateable(v)$, $q(n_v)$ cannot specify any relationship between $v′$ and v.

1-Learnability. Intuitively, 1-Learnable($n_v$) holds when $q(n_v)$ has a particular form of conditions in a where-clause. The property guarantees that $q(n_v)$ is learnable if $q(n_v) \in ancestors(n_v)$ and context(e) are given. Formally, 1-Learnable($n_v$) holds iff $q(n_v) = "for v in p where expr_w return v"$, where

1. $expr^*(v).path$ is a regular path expression that starts from the document root (Note that p does not have to start from the document root), and
2. $expr_w = \bigwedge_{v \in \{associateable(v) - associateable(v)\}} RS\{(v, v′)\}$, where $RS\{(v_1, v_2)\}$ is either True or $RS′\{(v_1, v_2)\}$. Here, $RS′\{(v_1, v_2)\}$ stands for one of the following expressions that satisfies a relationship between $v_1$ and $v_2$:

- (Rel1) $data(v_1) = data(v_2)$
- (Rel2) some $w$ in $v_1/q$ satisfies $RS′\{(v_1, v_2)\}$
- (Rel3) some $w$ in document()/$q$ satisfies $RS′\{(v_1, v_2)\}$

where q is a path expression with the child axis and optional position numbers or last() function (e.g., \(a[1]\) for \(b/c[last()]\)). The definition says that the condition in a where clause of $q(n_v)$ should be a conjunction of relationships between a variable $v_v$ and $v′ \in \{associateable(v_v) - associateable(v_v)\}$. It is also possible that two variables have an indirect association. Rel2 and Rel3 represent two patterns of indirect associations via $w$. Here, we refer to $w$ as a relay node of the relationship $RS′\{(v_1, v_2)\}$. Figure 10 illustrates indirect association between $c$ and $i$ through two relay nodes, where "some $ic in $i/incaregory/@category satisfies (some $ci in $c/@id satisfies ($ic=\$ci)\)" holds.

The point here is that, given an XML instance (e.g., Figure 10) and a pair of XML nodes ("H. Potter" and "book") in it, we can enumerate all the predicates that can appear in the where-clause of $q(n)$ if we know that 1-Learnable($n_v$) holds. Note that 1-Learnability’s definition implies that (1) only equality predicates and simple path expressions are contained in the where clause, both of which can be derived from the given XML instance in a straightforward way, and that (2) the number of possible predicates is finite.

Note that 0-Learnable($n_v$) is a special case where $expr^*(v).path = p$ and there is no condition in the where clause, so 0-Learnable($n_v$) ⇒ 1-Learnable($n_v$). In addition, if $n$ is the root, $depends(n) = \phi$ and $expr_w = True$. Therefore, 1-Learnable($n_v$)\&\&Root($n_v$) ⇒ 0-Learnable($n_v$).

[X1++$] An XQ-Tree t is in Class X1 iff t consists of the root node $n_1$ only and 1-Learnable($n_1$). Since 1-Learnable($n_v$)\&\&Root($n_v$) ⇒ 0-Learnable($n_v$), X1=X0.

[X1++] An XQ-Tree t is in X1+ iff for every node $n$ in $t$, 1-Learnable($n$).

[X1+++] An XQ-Tree t is in X1++ iff for every node $n$ in $t$, 1-Learnable($n$) \&\& 1-Learnable($n$), where 1-Learnable($n$) is defined in a similar way to 0-Learnable($n$).

The relationship among the classes is shown in Figure 11. The XQ-Tree in Figure 6 (ignoring the part in the box) is in X1++.

7 LEARN-X1++

1-Learnability’s definition and the XQuery’s semantics say that $q(n)$ may refer to variables defined in ancestors($n_v$). Therefore, $depends(n) = ancestors(n_v)$ when 1-Learnable($n_v$) holds. This suggests that algorithms to learn queries have to take care of traversal order of the XQ-Tree.

LEARN-X1++ works in the same way as LEARN-X0++ except that (1) it traverses the given XQ-Tree skeleton in
the depth-first order and (2) it incorporates LEARN-X1 instead of LEARN-X0 to learn the query fragment \( q(n) \) for each \( \text{XQ-Tree} \) node \( n \). LEARN-X1 learns \( q(n) \) when 1-Learnable(\( n \)) holds. It differs from LEARN-X0 in that (1) it takes as input \( context(n) \) and (2) its outputs (query fragments) contain where clauses.

7.1 LEARN-X1

We use the following notational convention: \( n_e \) is the \( \text{XQ-Tree} \) node whose query fragment \( q(n_e) \) is to be learned. \( e \) is the example object that has been dropped to \( n_e \). \( v_e \) is the variable that corresponds to \( e \) (which is assigned in a return clause in the \( \text{XQ-Tree} \) skeleton). \( context(e) = \{(v_{e1}, e_1), \ldots, (v_{em}, e_m)\} \). \( expr_f = \text{"in p"} \) and \( expr_w = \text{c} \). Then, \( q(n_e) = \{\text{for } v_e \text{ in p where } c \text{ return } v_e\} \) and \( EXT_e = \{\text{for } expr_f, v_e \text{ in p where } expr_w \land c \land (v_{e1} = e_1) \land \ldots \land (v_{em} = e_m) \text{ return } v_e\} \).

We represent the condition \( c \) in the where clause of \( q(n_e) \) as a set of predicates, and interpret \( c \) as the conjunction of the predicates in \( c \).

Note that if \( c \supset c' \), it is guaranteed that \( c \) is stronger than \( c' \) in the sense that \( \{\text{for } v \text{ in p where } c' \text{ return } v\} \) is contained in \( \{\text{for } v \text{ in p where } c \text{ return } v\} \).

Overview. Figure 12 illustrates the data flow among modules to implement LEARN-X1. The user has \( EXT_e \) in mind (Figure 12(a)). The algorithm learns \( EXT_e \) by receiving answers to (1) membership queries about \( EXT_e \) and (2) equivalence queries about the relationship between \( EXT_e \) and hypothesis \( EXT_e \). In order to compute \( EXT_e \) (Figure 12(b)), the algorithm must compute \( \hat{p} \) and \( \hat{c} \). P-Learner computes \( \hat{p} \). C-Learner (Section 7.2) takes as input \( context(e) \) and the Interaction History Table (\( IHT_e \)) to compute \( \hat{c} \).

Interaction History Table. \( IHT_e(Node, Ans, P, C) \) is a relation that records the user’s answer to learner’s queries for \( EXT_e \). Each tuple \( t \) in \( IHT_e \) records one answer (interaction) to a learner’s query. Attribute \( Node \) stores an XML node. Attribute \( Ans \) stores the answer to a learner’s query. For example, if \( Y \) is given to XLearner as an answer to a membership query on \( m \), then \( t.Node = m \) and \( t.Ans = Y \). It also records the reasons for the answer. If \( t.Ans = N \) because the path expression \( p \) does not accept \( t.Node, t.P = N \). If \( t.Ans = N \) because \( t.Node \) does not satisfy the condition \( c, t.C = N \). XLearner puts default values for \( t.P \) and \( t.C \) and corrects them if it finds inconsistencies in the user’s interactions [12]. If \( t.Ans = Y \), both \( t.P \) and \( t.C \) must be \( Y \). \( IHT_e \) always includes \( (e, Y, Y, Y) \) to record the fact that the dropped example node \( e \) is a positive example of \( EXT_e \).

7.2 C-Learner

C-Learner outputs the strongest condition consistent with all positive example nodes stored in \( IHT_e \). Because condition \( \hat{c} \) is a conjunction of predicates, we can map \( \hat{c} \) into a monotone (no negation) \( k \)-term \( x_1 \land \ldots \land x_k \), where each predicate in \( \hat{c} \) corresponds to a variable in the term. There is an algorithm to learn a monotone \( k \)-term that makes equivalence queries at most \( k \) times (Figure 13). In the algorithm, if \( i \) is included in \( A, x_i \) must be true. The algorithm always makes an equivalence query with the strongest condition so that every counterexample (an assignment to \( y_1, \ldots, y_k \)) is guaranteed to be a positive counterexample. Note that a counterexample can remove a number of unnecessary variables at one time.

We apply the algorithm to our context, using predicates instead of variables. The first issue to consider is how to derive the set of candidate predicates that can be included in \( \hat{c} \). Let an assignment \( a_0 \) be \( context(e) \cup \{(v_e, e)\} \). Intuitively, \( a_0 \) represents a situation where \( e \) is accepted as a member of \( EXT_e \). C-Learner computes the set of candidate predicates by computing the set of all predicates that hold about relationships between the dropped example \( e \) and all the XML nodes in \( context(e) \). Predicates used in 1-Learnable(\( n \)) are all considered. We use \( cond(context(e), (v_e, e)) \) to denote the set of candidate predicates. (How to compute the set is explained later.) Note that \( cond(context(e), (v_e, e)) \) is the “strongest” set of predicates for the first hypothesis (i.e., \( k = |cond(context(e), (v_e, e))| \) in the algorithm). Therefore, \( cond(context(e), (v_e, e)) \) may have unnecessary predicates that happen to hold only in \( a_0 \). When this occurs, C-Learner eliminates such predicates using counterexamples.

The second issue is what we should use as counterexamples in Figure 13 to compute \( \hat{c} \). As mentioned, C-Learner uses \( context(e) \) and \( IHT_e \) to compute \( \hat{c} \). Each assignment \( a_i \in \{context(e) \cup \{(v_e, t.Node)\} | t \in \sigma_{A_{ans}}(IHT_e) \} \) represents a situation where a positive example node for \( EXT_e \) is accepted. We use the set of predicates that hold in each situation as a counterexample. Formally, for each \( t \in \sigma_{A_{ans}}(IHT_e), cond(context(e), (v_e, t.Node)) \) serves as a counterexample. In the setting, the algorithm outputs the strongest condition consistent with all situations where the positive examples are chosen.

Computing \( cond(context(e), (v_e, e_i)) \). The computation is done with the aid of a \textit{data graph}. It is a graph structure of XML similar to the XQuery and XPath data model.
A = \{1, \ldots, k\}
while (true) {
    Let \( \hat{e} = \wedge_i A x_i \).
    Make an equivalence query with \( \hat{c} \).
    If the answer of the query is ‘OK’ (i.e., \( \hat{c} = c \)) {
        halt
    } else {
        let \( y_1, \ldots, y_k \) be the counterexample.
        \( A = A \cap \{ y_i = t_r \} \)
    }
}

Figure 13. Algorithm to Learn Monotone \( k \)-term from Positive Counterexamples

[17], where a tree structure represents an XML instance. In addition, the data graph has edges between nodes having the same value. We call such an edge a \( v \)-equality edge. Figure 10 is an example fragment of a data graph. C-Learner enumerates all the predicates that hold about relationships among example nodes. Note that conditions in where clauses are restricted to be a particular form by 1-Learnable(\( n \))’s definition, and the number of possible predicates is limited.

The main concern about the algorithm is management of the data graph. First, keeping all of the \( v \)-equality edges among nodes requires a large amount of additional data. Second, enumerating all paths between example nodes has an exponential complexity. We can use heuristics and known techniques to cope with the issues [12]. For example, we can take advantage of the fact that the values used for join conditions are limited, and can limit the maximal length of join paths. From our experience, the solution is adequate for typical XQuery queries.

8 Reducing the Number of Interactions

We denote by \( p, n \) and \( k \) a regular expression, the number of states in the minimal DFA representing \( p \), and the number of characters over which the language is defined. In XLearner, \( k \) corresponds to the number of XML element types. Let \( m \) be the length of the longest counterexample received. In our context, the length of a counterexample \( cc \) is the length of a tag sequence from the root to \( cc \). For the Angluin’s algorithm, the number of learner’s queries is in \( \tilde{O}(kmn^2) \). This means learning a simple regular path expression (corresponding to a DFA with a small number of states) could require hundreds of learner’s queries. The polynomial number of interactions is not sufficient for XLearner, because it is a real-world interactive system.

XLearner reduces the number of the learner’s queries by using properties specific to our context. Specifically, it uses the following two rules to automatically give default answers for the learner’s queries without asking the user for answers. As explained in Section 10, the rules dramatically reduce the number of learner queries.

R1: If Angluin’s algorithm makes a membership query on sequence \( s \) of tags and there is no \( m \) in the XML instance

\[ \mathrm{Ns}:: \quad \text{for } \ldots \text{return count(Ns.1)*10} \]
\[ \mathrm{Ns.1}:: \quad \text{for } $v$ in distinct(Ns.1.1) \text{return } $v$ \]

Figure 14. Nested Drop Box

s.t. \( \text{path}(m) = s \), \( N \) (No) is given to the algorithm. The current prototype uses the Relax NG for filtering, but other forms of metadata such as Graph Schema can be used as well.

R2: If a positive example \( m \) has been given to P-Learner, and the last tag of \( \text{path}(m) \) is \( \text{t1} \), \( N \) is given to the algorithm as answers to membership queries on \( s \) whose last tag is not \( \text{t1} \). If XLearner receives another positive counterexample with the last tag \( \text{t2}\neq \text{t1} \) for an equivalence query, XLearner backtracks and uses a new assumption that the last tag matches any kind of node. Finally, if XLearner receives a negative counterexample under the new assumption, XLearner discards all the assumptions and gives no more default answers. This is based on a heuristic that the last component of a typical path expression is likely to be a tag.

9 Expressive Power and an Extension

This section relaxes the assumption that the user is a minimally adequate teacher, which may be too strong for practical use. We extend LEARN-X1*+ by allowing it to receive the following three types of explicit specifications.

(1) Functions in Drop Boxes. So far, a Drop Box is assumed to be a place where the user drops an XML node taken from the XML Browser. In general, however, mapping of an XML schema to another sometimes involves functions, such as aggregation functions. We now introduce the concept of a Nested Drop Box and allow the user to explicitly specify functions. If the user types the name of functions in a Drop Box, XLearner opens a new (nested) Drop Box corresponding to each parameter of the function (Figure 14 (left)). According to the type of return values and parameters of a given function, XLearner rewrites the structure of the XQ-Tree (Figure 14 (right)), allows the user to drop example nodes in the Drop Boxes, and learns the query fragments from simple yes-or-no interactions.

(2) Keys for Sorting. We allow XLearner to receive keys for sorting through a dialog box called an OrderBy Box. An OrderBy Box appears when the user clicks on an element to be sorted. After receiving the sort key(s), XLearner inserts an order by clause into the query fragment.

(3) Explicit Selection/Join Conditions. This is specified in a Condition Box, as explained in Section 2. Note that in contrast to the other types of explicit specifications, XLearner automatically learns a certain kind of join condition. The question, then, is how to integrate the learning schema and the Condition Box. Interestingly, XLearner can determine
when a Condition Box should be used by checking the consistency in $IHT_c$ [12].

There are two types of Condition Boxes: a Positive Condition Box (PCB) and a Negative Condition Box (NCB). A PCB is used to specify why the dropped positive example is included in the extent, while an NCB is used to specify why the negative counter example is not included in the extent. The user can choose either type. The one used in Section 2 is a PCB. If the user specifies a condition $c$ with an NCB, condition $\neg c$ is used as a condition. NCB is especially useful when the user wants to use empty predicate\(^4\) in a condition, because there is no positive example node that can be used as its parameter. The subexpression in the box in Figure 6 can be created by this extension.

With this extension of the basic framework, XLearner has a practical expressive power. Let $X^{1*+E}$ be a class of XQ-trees learnable by LEARN-X1*+ with the extension. Then, we define a set $XQ_t$ of XQuery queries as follows: Given an XML instance $I$, an XQuery query $Q$ is in $XQ_t$ if there exists a query $Q'$ s.t. $Q'(I) = Q(I)$ and the XQ-tree representation of $Q'$ is in $X^{1*+E}$. The example query in Figure 2 (including the box part) is in $XQ_t$. Note that $XQ_t$ is parameterized by instance $I$; it is natural for the set to depend on $I$ because we discuss a system that receives real examples from an XML instance.

Figure 15 shows the percentages of queries in XMark and XML Use Cases that are included in $XQ_t$. Specifically, they include 19 out of the 20 XMark queries\(^5\), and 11 out of 12 XMLQuery Use Case “XMP” (Experiences and Exemplars) queries. The main reason for 0% in Use Case “NS” is that the queries contain special matching patterns that use namespaces. The queries in Use Case “PARTS” contain recursive user-defined functions. The queries in Use Case “STRONG” exploit information on strongly typed data.

Figure 15. Expressive Power of XLearner

<table>
<thead>
<tr>
<th>Name</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>XMark</td>
<td>95% (19/20)</td>
</tr>
<tr>
<td>UC “XMP”</td>
<td>91.9% (11/12)</td>
</tr>
<tr>
<td>UC “TREE”</td>
<td>83.3% (58/55)</td>
</tr>
<tr>
<td>UC “SEC”</td>
<td>69% (7/11)</td>
</tr>
<tr>
<td>UC “R”</td>
<td>77.8% (14/18)</td>
</tr>
<tr>
<td>UC SGML</td>
<td>100% (11/11)</td>
</tr>
<tr>
<td>UC “STRONG”</td>
<td>30% (2/6)</td>
</tr>
<tr>
<td>UC “NS”</td>
<td>0% (0/8)</td>
</tr>
<tr>
<td>UC “PARTS”</td>
<td>0% (0/1)</td>
</tr>
<tr>
<td>UC “STRONG”</td>
<td>0% (0/2)</td>
</tr>
</tbody>
</table>

10 Experiments

We counted the number of interactions required for learning queries and evaluated the effect of the mechanisms

\(^4\)The empty predicate returns true if the parameter is an empty sequence.

\(^5\)For Q18 that uses a user-defined function, XLearner learned an equivalent, but different query without any user-defined function.

In Figure 16, D&D is the number of dropped example nodes. Note that the user is allowed to specify an arbitrary function in each Drop Box. Therefore, we need a measure of its complexity. We use the number of terminal nodes in the function’s abstract syntax tree. Terminal nodes include function names, values, and dropped example nodes. For example, the number of terminal nodes in multiply(plus(30, 40), 2) equals 5. \#t in the parenthesis is the number of terminal nodes. MQ is the number of membership queries, and CE is the number of counterexamples given by the user. CB is the number of Condition Boxes that are invoked to specify explicit selection/join conditions. We give the numbers of terminal nodes again.

<table>
<thead>
<tr>
<th>Name</th>
<th>Reduced</th>
<th>MQ</th>
<th>OB</th>
<th>CB/#t</th>
<th>Reduced</th>
<th>R1,R2,Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>1/1</td>
<td>5</td>
<td>1</td>
<td>1(3)</td>
<td>0</td>
<td>2434(2412, 486, 464)</td>
</tr>
<tr>
<td>Q2</td>
<td>1/1</td>
<td>1</td>
<td>1</td>
<td>1(4)</td>
<td>0</td>
<td>2439(2412, 486, 463)</td>
</tr>
<tr>
<td>Q3</td>
<td>2/2</td>
<td>0</td>
<td>1</td>
<td>1(3)</td>
<td>0</td>
<td>4878(4832, 972, 926)</td>
</tr>
<tr>
<td>Q4</td>
<td>1/1</td>
<td>0</td>
<td>1</td>
<td>1(9)</td>
<td>0</td>
<td>1627(1608, 405, 380)</td>
</tr>
<tr>
<td>Q5</td>
<td>2/2</td>
<td>0</td>
<td>1</td>
<td>1(5)</td>
<td>0</td>
<td>1627(1612, 405, 380)</td>
</tr>
<tr>
<td>Q7</td>
<td>3/8</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>77.8%(14/18)</td>
</tr>
<tr>
<td>Q8</td>
<td>2/3</td>
<td>0</td>
<td>1</td>
<td>1(5)</td>
<td>0</td>
<td>2604(7557, 7557, 7557)</td>
</tr>
<tr>
<td>Q9</td>
<td>2/2</td>
<td>0</td>
<td>1</td>
<td>1(3)</td>
<td>0</td>
<td>4051(4023, 881, 853)</td>
</tr>
<tr>
<td>Q10</td>
<td>1/2</td>
<td>0</td>
<td>1</td>
<td>1(3)</td>
<td>0</td>
<td>2694(26756, 5589, 5351)</td>
</tr>
<tr>
<td>Q11</td>
<td>2/3</td>
<td>0</td>
<td>1</td>
<td>1(5)</td>
<td>0</td>
<td>4066(4025, 891, 850)</td>
</tr>
<tr>
<td>Q12</td>
<td>2/3</td>
<td>0</td>
<td>1</td>
<td>1(8)</td>
<td>0</td>
<td>4066(4025, 891, 850)</td>
</tr>
<tr>
<td>Q13</td>
<td>2/2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4808(4822, 972, 926)</td>
</tr>
<tr>
<td>Q14</td>
<td>1/1</td>
<td>5</td>
<td>1</td>
<td>1(3)</td>
<td>0</td>
<td>2426(2408, 486, 464)</td>
</tr>
<tr>
<td>Q15</td>
<td>1/1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12637(12564, 1053, 1020)</td>
</tr>
<tr>
<td>Q16</td>
<td>1/1</td>
<td>1</td>
<td>1</td>
<td>1(2)</td>
<td>0</td>
<td>2438(2422, 486, 470)</td>
</tr>
<tr>
<td>Q17</td>
<td>1/1</td>
<td>0</td>
<td>1</td>
<td>1(2)</td>
<td>0</td>
<td>1171(1161, 405, 389)</td>
</tr>
<tr>
<td>Q18</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1627(1608, 405, 386)</td>
</tr>
<tr>
<td>Q19</td>
<td>2/2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4848(4804, 972, 926)</td>
</tr>
<tr>
<td>Q20</td>
<td>3/8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4850(4824, 972, 926)</td>
</tr>
</tbody>
</table>

Figure 16. The Number of Interactions for Learning

106(98, 32, 24)
ependent of the interactions; they are the number of element types defined in the DTD (MQ is affected. See Section 8.), the size of XQ-tree (all measurements), the number of states in the minimum dfa corresponding to each path regular expression (MQ, CE), the density of the data graph (affecting the number of predicates enumerated by C-Learner, thus CE), and the regularity of the data structure (Reduced). Note that the size of the data graph is not included in the factors. Dynamic factors mean which XML nodes the user gives as counterexamples against equivalence queries (MQ, CE, Reduced) and which XML nodes to choose for membership queries (CE). Measurements shown in the figure are basically the “best-case” measurements with the XML instances. If the “worst-case” measurement is worse than the “best-case,” the “worst-case” is shown in square brackets.6

Following is a key observation: Rules R1 and R2 dramatically reduce the number of interactions, thereby showing that our approach can be practical. R1 suppresses membership queries if they are inconsistent with the XML instance. Therefore, the more regular the structure of the XML instance, the fewer membership queries XLearner will ask. The structures of data in XMark and XML Use Cases are relatively regular, so the rule eliminates a great number of interactions. The regularity also accounts for the reason why there is no difference between the best and worst cases in MQ and Reduced, which are affected by the structure of paths from the document root to examples.

Another important observation is that, in practice, it does not matter so much which example is chosen as a counterexample. The reason is that the data graph is relatively sparse and the performance of C-Learner is little affected by the choice of a counterexample. This also explains why CE is always 0 or a small number and is not a function of D&D (the size of the XQ-Tree skeleton) in the results. Theoretically, MQ scales linearly with the number of D&D. But because experimental results show XLearner requires 0 to only a few interactions for learning each typical query fragment, we can say that XLearner requires a relatively small number of interactions for a general XQuery query, even if it is larger than the queries used in the experiments. For example, Figure 16 shows that learning Q10 consisting of 12 XQuery nodes requires at most three interactions after the user drops and drops example nodes.

11 Discussion and Conclusion

XLearner is more appropriate for restructuring than retrieval. Using real examples is very intuitive; we developed a Web-site construction tool in a similar approach and demonstrated that even users with no knowledge of database-related concepts could specify fairly complex data manipulations [11]. The user, however, must always find examples. We believe the proposed approach is complementary to other retrieval-oriented approaches and we can combine other approaches and XLearner in appropriate ways. One possible extension is to incorporate known search mechanisms into XLearner to find examples that satisfy given conditions. Another direction is to allow the user to specify “artificial” examples in the spirit of QBE. Details of such extensions are beyond the scope of this paper.

In this paper, we explained XLearner, a tool for developing XML mapping queries using machine learning techniques. Specifically, we presented algorithms for learning several classes of XQuery queries. We also proposed some extensions and showed that our most elaborate algorithm has a practical expressive power and can learn a large set of XQuery queries. Beyond that, we presented experimental results where the number of required interactions is small. Future work includes development of a mechanism to reuse past interactive operations.

Acknowledgments

We would like to thank Akiyoshi Nakamizo for participating in the implementation. The research has been supported in part by the Grant-in-Aid for Scientific Research from JSPS and MEXT.

References


---

6We examined the XML instances and selected appropriate example nodes by hand for those cases.