Rewriting Regular XPath Queries on XML Views

Wenfei Fan 1,2, Floris Geerts 1,3, Xibei Jia 1, Anastasios Kementsietsidis 1
1 University of Edinburgh 2 Bell Laboratories 3 Hasselt University/Transnational Univ. of Limburg
{wenfei, fgeerts, xjia, akements}@inf.ed.ac.uk

Abstract

We study how to rewrite XML queries posed on virtual XML views into equivalent queries on the underlying XML document. For XML queries we start with a fragment of XPath, which supports recursion (the descendant-or-self axis ‘//’), union and complex filters (predicates). This class of XPath queries is commonly used in practice and is essential to XQuery, XSLT and XML Schema. We consider XML views defined by annotating a view DTD with a collection of (regular) XPath expressions, along the same lines as how commercial systems specify XML views [15, 21, 20]. An XML view defined as above is a mapping $\sigma: D \rightarrow D_V$, computes an XPath query $Q'$ over the document DTD $D_V$ such that for any XML tree $T$ of $D$, $Q(\sigma(T)) = Q'(T)$.

1. Introduction

In many applications users are allowed to access an XML document only by querying a view of the data. The need for this is evident in, for example, enforcing access control on XML data [2, 5, 9]. To prevent improper disclosure of sensitive or confidential information of XML data residing in a server, the server defines an XML view for each group of users, consisting of all and only the information that the users are authorized to access. While the users may query the view, they are not allowed to directly query or access the underlying document (referred to as the source). With this comes the need to answer queries posed on the views. One way to do this is to first materialize the views and then directly evaluate queries on the views. However, it is often too costly to materialize and maintain a large number of views, a common scenario when many groups of users with different access privileges query the same source. A more realistic approach is to rewrite (aka. translate, reformulate) queries on the views into equivalent queries on the source, and return the rewritten queries on the source without materializing the views, and return the answers to the users.

We study the problem of answering queries posed on virtual views of XML documents, a problem commonly encountered when enforcing XML access control and integrating data. We approach the problem by rewriting queries on views into equivalent queries on the underlying document, and thus avoid the overhead of view materialization and maintenance. We consider possibly recursively defined views and study the rewriting of both XPath and regular XPath queries. We show that while rewriting is not always possible for XPath over recursive views, it is for regular XPath; however, the rewritten query may be of exponential size. To avoid this prohibitive cost we propose a rewriting algorithm that characterizes rewritten queries as a new form of automata, and an efficient algorithm to evaluate the automaton-represented queries. These allow us to answer queries on views in linear time. We have fully implemented a prototype system, SMOQ, which yields the first regular XPath engine and a practical solution for answering queries over possibly recursively defined XML views.

Example 1.1: Consider a hospital DTD $D$ shown in Fig. 1(a). A hospital document of $D$ consists of a list of departments, and each department has a list of in-patients (i.e. patients who are currently residing in the hospital; we...
use ‘*’ on an edge to indicate a list). For each patient, the hospital maintains her name (pname), address, records of visits, each including the visit date and treatment which is either a test or some medication (dashed edges indicate disjunction), as well as information about the treating doctor. Each name, pname, street, city, zip, date, type, dname, specialty has a single text node (PCDATA) as its child (omitted in the figure). The hospital also maintains family medical history by means of the recursively defined parent and sibling. It records the same information of ancestors with those of in-patients, by sharing the description for patients.

A view \( \sigma_0 \) is defined for a research institute studying inherited patterns of heart disease, with the view DTD depicted in Fig. 1(b) (the view is defined in Example 2.2). Obligated by the Patient Privacy Act, the view reveals only those patients who have heart disease, along with their parent hierarchy. While the institute may access diagnosis information of those patients and their ancestors, it is denied access to their name, address, test and doctor data.

Consider an XPath query \( Q \) posed on the view, which is to find patients whose ancestors also had heart disease:

\[
Q: \text{patient}[\text{/record/diagnosis/text()} = \text{‘heart disease’}]\]

Here * denotes a wildcard, i.e., any element. However, it is impossible to rewrite \( Q \) on the view to an equivalent query (in the XPath fragment mentioned above) on the underlying hospital document. This is because ‘//’ in \( Q \) is supposed to traverse only the parent hierarchy on the view, i.e., a sequence of the (parent/patient) pattern; however, when translated to a query \( Q' \) on the source, \( Q' \) necessarily retains ‘//’ since the view DTD is recursive, and ‘//’ in \( Q' \) may access siblings of those patients, although siblings are not in the view and are not allowed to be accessed. An incorrect translation may lead to serious security breach.

In response to this we develop both fundamental results and practical techniques for the rewriting problem. The main contributions of the paper include the following.

1. **Closure Properties.** On the theoretical side, we study the closure property of XPath under query rewriting: is it always possible to rewrite XPath queries on views to XPath queries on the source? We prove that XPath is not closed under query rewriting for recursive views. In light of this we consider a mild extension of XPath, regular XPath [19], which uses the general Kleene closure \( E^* \) instead of the ‘//’ axis. We show that regular XPath is closed under rewriting for arbitrary views, recursive or not. Since regular XPath subsumes XPath, any XPath queries on views can be rewritten to equivalent regular XPath queries on the source.

However, the rewriting problem is EXPTIME-complete: for a (regular) XPath query \( Q \) over even a (non-)recursive view, the rewritten regular XPath query on the source may be inherently exponential in the size of \( Q \) and the view DTD \( D_V \). This tells us that rewriting is beyond reach in practice if \( Q \) is directly rewritten into regular XPath.

On the practical side, to avoid the exponential blow-up we develop the following techniques for answering (regular) XPath queries posed on XML views.

2. **Automaton-based rewriting for (regular) XPath.** We introduce a rewriting method based on a notion of mixed finite state automata (MFA) to represent rewritten regular XPath queries. An MFA is a nondeterministic finite automaton (NFA) “annotated” with alternating finite state automata (AFA), which characterize data-selection paths and filters of a regular XPath query \( Q \), respectively. The algorithm rewrites \( Q \) into an equivalent MFA \( \mathcal{M} \). In contrast to the exponential blowup, the size of \( \mathcal{M} \) is bounded by \( O(|Q||\sigma| |D_V|) \). This makes it possible to answer queries on views via rewriting. To our knowledge, although a number of automaton formalisms were proposed for XPath and XML stream (e.g. [6, 13]), they cannot characterize regular XPath queries, as opposed to MFA.

3. **Evaluation of rewritten query.** We provide an efficient algorithm for evaluating MFA \( \mathcal{M} \) (rewritten regular XPath queries) on XML source \( T \). While there have been a number of evaluation algorithms developed for XPath, none is capable of processing regular XPath queries. Previous algorithms for XPath (e.g., [16]) require at least two passes of \( T \): a bottom-up traversal of \( T \) to evaluate filters, followed by a top-down pass of \( T \) to select nodes in the query answer. In contrast, our evaluation algorithm combines the two passes into a single top-down pass of \( T \) during which it both evaluates filters and identifies potential answer nodes. The key idea is to use an auxiliary graph, often far smaller than \( T \), to store potential answer nodes. Then, a single traversal of the graph suffices to find the actual answer nodes. The
algorithm effectively avoids unnecessary processing of sub-trees of $T$ that do not contribute to the query answer. It is not only the first efficient algorithm for evaluating regular XPath queries (MFA), but also provides an efficient (alternative) algorithm to evaluate XPath queries.

4. Implementation and experimental study. We have implemented a prototype system SMOQE (Secure MODular Query Engine [10]) for answering queries on XML views, fully supporting the rewriting and evaluation techniques mentioned above. Using SMOQE we have conducted an experimental study, which clearly demonstrates that our evaluation techniques are efficient and scale well. For regular XPath queries, we compared the SMOQE evaluation of queries with that of their XQuery translation, and found that the latter requires considerably more time. Furthermore, SMOQE outperforms the widely used XPath engine Xalan (default XPath implementation in Java 5), whether Xalan uses its interpretive processor or its high performance compiling processor (XSLT TC), when evaluating XPath queries.

In summary, we provide the first practical and complete solution to answering regular XPath queries posed on (virtual and possibly recursively defined) XML views. It is provably efficient: it has a linear-time data complexity and a quadratic combined complexity. Furthermore it yields the first efficient technique for processing regular XPath queries, whose need is evident since regular XPath is increasingly being used both as a stand-alone query language and as an intermediate language in query translation [11].

Organization. Section 2 reviews (regular) XPath and XML views. Section 3 discusses the closure property of (regular) XPath rewriting. Section 4 introduces MFA and Section 5 describes the rewriting algorithm. Section 6 presents the MFA evaluation algorithm, followed by experimental results in Section 7. Related work is discussed in Section 8, followed by conclusions in Section 9.

2. Background

In this section we review XPath [4], regular XPath [19], DTDs and XML views considered in this paper.

2.1. XPath and Regular XPath

We consider a class of regular XPath queries proposed and studied in [19], denoted by $\mathcal{X}_{\text{reg}}$ and defined as follows:

$$Q ::= \varepsilon \mid A \mid Q/Q \mid Q \cup Q \mid Q^* \mid Q[Q],$$

$$q ::= Q \mid Q/\text{text}() = 'c' \mid \neg Q \mid Q \land Q \mid Q \lor Q$$

where $\varepsilon$ is the empty path (self), $A$ is a label (tag), ‘$\cup$’ represents union, ‘$\land$’ is the child-axis, and $*$ is the Kleene star; $[Q]$ is referred to as a filter, in which $Q$ is an $\mathcal{X}_{\text{reg}}$ expression, $c$ is a string constant, and $\neg, \land, \lor$ are the Boolean negation, conjunction and disjunction, respectively. Regular XPath extends regular expressions by allowing filters [19], and extends XPath by supporting Kleene closure $Q^*$ as opposed to the restricted recursion ‘$\prime$’ (the descendant-or-self axis).

Like XPath queries, when an $\mathcal{X}_{\text{reg}}$ query $Q$ is evaluated at a node $v$ in an XML tree $T$, it returns the set of nodes of $T$ reachable via $Q$ from $v$, denoted by $v[Q]$.

We also study an XPath fragment of $\mathcal{X}_{\text{reg}}$, denoted by $\mathcal{X}$, which is defined by replacing $Q^*$ with ‘$\prime$’ in the definition above. Note that given a DTD $D$ of the documents on which queries are posed, ‘$\prime$’ is expressible in $\mathcal{X}_{\text{reg}}$ as $\left(\bigcup E\right)^*$, where $\bigcup E$ denotes the union of all the labels in $D$.

Example 2.1: Consider an XML document $T$ conforming to the document DTD $D$ in Fig. 1(a). The regular XPath query $Q = \text{department/patient[} q_0 \land (q_1/(q_1)^*))]/\text{pname}$

$q_0 = \text{visit/treatment/medication/diagnosis/text() = “heart disease”}$

$q_1 = \text{parent/patient}[\neg q_0]/\text{parent/patient[}q_0]$ when evaluated on $T$, returns the names of patients who have heart disease and the disease appears in their ancestors but always skips a generation. Such queries, which look for certain patterns, are often encountered in medical research. Note that the query is in the fragment $\mathcal{X}_{\text{reg}}$, but is not expressible in the XPath fragment $\mathcal{X}$.

In this work we focus on regular XPath queries with only downward modalities since they are most commonly used in practice. As will be seen shortly, rewriting queries is already challenging in this setting. It is thus necessary to understand rewriting of these basic queries before dealing with full-fledged XPath or XQuery.

2.2. DTD

Following [9], we represent a DTD $D$ as a triple $(E\ell, P, r)$, where $E\ell$ is a finite set of element types; $r$ is a distinguished type in $E\ell$, called the root type; $P$ defines the element types: for each $A$ in $E\ell$, $P(A)$ is a regular expression of the form: $\text{str, } c, B_1, \ldots, B_n, \text{ or } B_1 + \ldots + B_n$. Here $\text{str}$ denotes PCDATA, $c$ is the empty word, $B_i$ is either $B$ or of the form $B^*$ where $B$ is in $E\ell$ (referred to as a child type of $A$), and ‘$+$’, ‘$'$’ and ‘$*$’ denote disjunction (with $n > 1$), concatenation and the Kleene star, respectively. We refer to $A \rightarrow P(A)$ as the production of $A$. This form of DTD’s does not lose generality since any DTD can be converted to a DTD of this form by using new element types.

A DTD can be represented as a graph, as shown in Fig. 1. It is recursive if the corresponding graph is cyclic. For example, both DTD’s depicted in Fig. 1 are recursive.

2.3. XML Views

We consider views defined by annotating a DTD [9]. This is similar in spirit to XML view specification in commercial systems, e.g. annotated XSD’s (AXSD) in Oracle XML DB [21] and Microsoft SQL Server 2000 SQLXML [20], and Document Access Definitions (DAD) of IBM DB2 XML Extender [15]. Specifically, we define an XML view as a mapping $\sigma : D \rightarrow D_V$, where $D$ is a document DTD, $D_V$ is a view DTD. Given an XML document $T$ of $D$, the mapping generates an XML view $\sigma(T)$ that conforms to the view...
<table>
<thead>
<tr>
<th>Query rewriting</th>
<th>Views</th>
<th>Closure</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>from X to X'</td>
<td>non-rec.</td>
<td>Yes [9]</td>
<td>EXPTIME-complete</td>
</tr>
<tr>
<td>from A to X</td>
<td>recursive</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>from X to X'</td>
<td>arbitrary</td>
<td>Yes</td>
<td>EXPTIME-complete</td>
</tr>
<tr>
<td>from X to Xreg</td>
<td>arbitrary</td>
<td>Yes</td>
<td>EXPTIME-complete</td>
</tr>
</tbody>
</table>

**Figure 2. Closure property and complexity**

DTD $D_V$. More specifically, for each element type $A$ and its child type $B$ in $D_V$ (i.e., each edge $(A, B)$ in the DTD graph of $D_V$), $\sigma$ maps $(A, B)$ to a query $\sigma(A, B)$ defined on documents $T$ of $D$. Intuitively, given an $A$ element, $\sigma(A, B)$ generates its $B$ children in the view by extracting data from $T$. The query $\sigma(A, B)$ is in the regular XPath fragment $X_{reg}$ given above. The XML view is recursive if the view DTD $D_V$ is recursive.

**Example 2.2:** Figure 1(c) defines the view $\sigma_0$ described in Example 1.1. The semantics of $\sigma_0$, informally presented, is as follows: Given a hospital document $T$, $\sigma_0$ generates a view $\sigma_0(T)$ top-down, which conforms to the view DTD of Fig. 1(b). The query $Q_1$ (i.e., $\sigma_0(hospital, patient)$) extracts from $T$ those patients who have heart disease. For the patients extracted by $Q_1$, (a) $Q_2$ finds their parent nodes, which are in turn processed by $Q_3$ and then inductively by $Q_2$ and $Q_3$ to form the parent hierarchy, and (b) $Q_3$ finds the record (i.e., visit) data, which can be either be empty (i.e., test) or diagnosis, handled by $Q_5$, $Q_6$, respectively. □

**3. The Closure Property of (Regular) XPath**

We next study the closure property and complexity of XPath and regular XPath query rewriting. The main results of this section are summarized in Fig. 2.

Formally, an XML query language $L$ is closed under rewriting if there exists a computable function $F : L \rightarrow L$ that, given any view definition $\sigma : D \rightarrow D_V$ and any query $Q$ in $L$ over $D_V$, computes query $Q' = F(Q)$ in $L$ such that for any document $T$ of $D$, $Q(\sigma(T)) = Q'(T)$. While one may consider translating an XPath query $Q$ to an equivalent $Q'$ in a richer language, e.g., XQuery or XSLT, it is vastly preferable to have an XPath translation since it is more efficient to evaluate XPath queries than queries in the aforementioned Turing-complete languages. The closure property is desirable since rewriting should not be penalized by paying the higher price for evaluating and optimizing queries in a richer language than that of the original query.

It was shown in [9] that the class $X$ of XPath queries defined in Section 2 is closed under query rewriting for non-recursive views. However, below we show that in the presence of recursion in a view definition, this is no longer the case (even when the annotating queries are in $X$).

**Theorem 3.1:** For recursively defined XML views, the fragment $X$ is not closed under query rewriting. □

In contrast, the fragment $X_{reg}$ of regular XPath given in the last section is closed under query rewriting:

**Theorem 3.2:** For arbitrary XML views (recursive or non-recursive), $X_{reg}$ is closed under rewriting. □

**Example 3.1:** Recall the view $\sigma_0 : D \rightarrow D_V$ defined in Example 2.2 and the query $Q$ given in Example 1.1. Using the queries $Q_1$, $Q_2$, $Q_3$, $Q_4$ and $Q_6$ from the view specification in Fig. 1(c), we can compute a correct rewriting $Q'$ of query $Q$. Specifically: $Q' = Q_1(Q_2/Q_4)/(Q_2/Q_4)^*/Q_3/Q_6/text() = \text{‘heart disease’}$. For any document $T$ that conforms to $D$, $Q'(T) = Q(\sigma_0(T))$. □

Although it is always possible to rewrite a (regular) XPath query on a view to an equivalent regular XPath query on the source, it is often prohibitively expensive if it is to directly compute $X_{reg}$ queries as output. Indeed, the rewriting problem subsumes the problem for translation from NFA's to regular expressions. The latter problem is EXPTIME-complete [8]; the size of the explicit representation of a regular expression is exponential in the size of the NFA. Worse still, it remains exponential even if the NFA is acyclic.

**Corollary 3.3:** There exist a view definition $\sigma : D \rightarrow D_V$ and a query $Q$ in $X$ such that for any $Q'$ in $X_{reg}$, if $Q(\sigma(T)) = Q'(T)$ for all XML trees $T$ of $D$, then the size $|Q'|$ of $Q'$, when represented as an $X_{reg}$ query, is exponential in $|Q|$ and the size $|D_V|$ of $D_V$. The lower bound remains intact even when $D_V$ is non-recursive. □

**4. Mixed Finite State Automata**

The exponential lower bound of Corollary 3.3 tells us that a direct rewriting into (regular) XPath is beyond reach in practice. To overcome this, in this section we introduce a new representation of $X_{reg}$ queries, referred to as mixed finite state automata (MFA). Along the same lines as NFA for regular expressions, MFA characterize $X_{reg}$ queries and avoid the exponential blowup of rewriting. Leveraging MFA we shall present a practical solution to the rewriting problem by providing (a) a low polynomial-time algorithm for rewriting $X_{reg}$ queries on a view into the MFA-presentation of equivalent $X_{reg}$ queries on the source (Section 5), and (b) a linear-time algorithm for directly evaluating the MFA-presentation of $X_{reg}$ queries on the source (Section 6).

While a regular expression can be efficiently represented as a graph or a NFA, for $X_{reg}$ queries a notion of automaton representation is not yet available. The difficulties of characterizing an $X_{reg}$ query $Q$ as an automaton include the following: (a) $Q$ typically involves both “selecting” paths that are to extract and return nodes, and filters that constrain the extraction; (b) a filter [q] in $Q$ may involve Boolean operators ‘\&\&’, ‘\|’, ‘\,', and constant test $p/text() = \text{‘c’}$, which are not encountered in regular expressions; (c) worse still, it may be nested: $q$ itself may be a query of the form $p[q_1]$; and (d) the sub-query $p$ of $p^*$ may itself contain Kleene closure.

Mixed finite state automata (MFA). In light of this we define an MFA $M$ as a nondeterministic finite automaton...
(NFA) in which a state may be annotated with an alternating finite state automaton (AFA). Intuitively, the NFA in $M$ is to capture the selecting paths of an $X_{eq}$ query $Q$ and the AFA’s to characterize the filters in $Q$.

Formally, an AFA $M$ is defined to be $(N_s, \lambda)$, where (a) $\lambda$ is a set of bindings $X_i = A_{i}^{\text{fa}}$, $X_i$ is a name and $A_{i}^{\text{fa}}$ is an AFA as defined below; (b) $N_s = (K_s, \Sigma_s, \delta_s, s, F, \lambda)$ is a variation of NFA, referred to as the selecting NFA of $M$, where $K_s, \Sigma_s, \delta_s, s, F$ are the states, alphabet, transition function, start state and final states as in the standard NFA definition; and $\lambda$ is a partial mapping from $K_s$ to names $X_i$, i.e., a state in $N_s$ may be annotated with a single $X_i$.

We employ a variation of AFA’s [24] to represent $X_{eq}$ filters. We define an AFA $A_{i}^{\text{fa}}$ to be $(K, \Sigma, \delta, s, F, e)$, where (a) $K$ is a set of states partitioned into $K_{op}$, $K_1$ and $K_2$, where $K_{op}$ is a set of operator states marked with AND, OR or NOT, $K_1$ is a set of transition states, and $F$ is a set of final states optionally annotated with predicates of the form $\text{text}(= \neq \text{position})=k$; (b) $\Sigma$ is a set of labels; (c) $s$ is the start state in $K$; and (d) $\delta$ is the transition function defined as follows. (1) For a state $s_1$ in $K_{op}$, $\delta$ is only defined for empty string $\epsilon$ and $\delta(s_1, \epsilon) = K'$, where $K'$ is a subset of $K$. In particular, if $s_1$ is marked with NOT, $K'$ has a single state in it. (2) For each state $s_2$ in $K_1$, $\delta$ is only defined for a single label $A \in \Sigma$ and $\delta(s_2, A)$ contains a single state in $K$. (3) $\delta$ is not defined for any state in $F$. Observe that except for operator states marked with AND or OR, from each state at most one state can be reached via $\delta$. These operator states capture Boolean operators $\wedge, \lor$ and $\neg$ in $X_{eq}$ filters.

**Example 4.1:** Consider an $X_{eq}$ query $Q_0$ posed on an XML tree conforming to the DTD of Fig. 1(b), which is to find all patients who have an ancestor diagnosed with heart disease:

$$Q_0 = (\text{patient}/\text{parent})/\text{patient}[q_0],$$

$$q_0 = (\text{parent}/\text{patient})/\text{record}/\text{diagnosis}[\text{text}]=\text{“heart disease”}.$$

Consider MFA $M_0$ in Fig. 3. It consists of a selecting NFA $N_s$ (shown at the top of the figure), and an AFA $A_{0}^{\text{fa}}$, corresponding to the filter $q_0$ (shown at the bottom). The MFA $M_0$ is equivalent to $Q_0$, in the sense that when evaluating $M_0$ at a node $n$ in an XML tree $T$ (described below), it returns the same set $n[M_0]$ of nodes as $n[Q_0]$.

The (conceptual) evaluation of $M_0$ is illustrated, by example, in Fig. 4. At the root node 1 of the tree, $M_0$ associates a set $\{s_1, s_3\}$ of $N_s$ states, where $s_1$ is the start state of $N_s$ and $s_3$ is reached from $s_1$ via an $e$-transition. It then inspects the children of node 1: for all its children labeled patient (nodes 2 and 9), it associates them with states $s_2, s_4$, moves down to these children and processes them inductively, in parallel. At a node associated with state $s_2$, for all its children labeled record (nodes 6 and 8), if one child is labeled diagnosis and carrying text “heart disease”, and if so, $X_2(s_6)$ is assigned true as well. Similarly, $X_2(s_4)$ is computed and becomes true if it has a descendant that is reachable via (parent/patient)/record/diagnosis and carries text “heart disease”. If either $X_2(s_2)$ or $X_2(s_4)$ is true, then $X_2(s_1)$ is true and so is the output $2[A_{0}^{\text{fa}}]$. This is not the case here, however, and $A_{0}^{\text{fa}}$ returns false.

Observe the following. (a) Although $A_{0}^{\text{fa}}$ traverses the subtree top-down, the Boolean variables are computed bottom-up. (b) In $A_{0}^{\text{fa}}$ the only operator states are OR states $(s_{A2}, s_{A4})$; but AND and NOT states can be processed similarly. (c) The conceptual evaluation requires multiple passes over a subtree, one pass for each filter. In contrast, our evaluation algorithm in Section 6 requires only one pass of the input tree, regardless of the number of filters.
Equivalence of MFA and $\mathcal{X}_{\text{reg}}$ queries. An MFA $\mathcal{M}$ and an $\mathcal{X}_{\text{reg}}$ query $Q$ are equivalent if for each XML tree $T$ and any node $n$ in $T$, $n[\mathcal{M}] = n(Q)$, where $n[\mathcal{M}]$ (resp. $n(Q)$) denotes the result of evaluating an MFA $\mathcal{M}$ (resp. $Q$) at $n$.

The result below tells us that we can identify a class of MFA’s, namely, MFA’s with a syntactic restriction on AFA’s called the split property, to precisely capture the fragment $\mathcal{X}_{\text{reg}}$ of regular XPath queries; as a result, MFA’s can be used to represent $\mathcal{X}_{\text{reg}}$ queries.

Theorem 4.1: For any $\mathcal{X}_{\text{reg}}$ query $Q$, there exists an equivalent MFA $\mathcal{M}$ with the split property, and vice versa. 

5. Rewriting Algorithm

We now present an efficient algorithm, called rewrite (not shown due to space constraints), for rewriting (regular) XPath queries on arbitrary views into equivalent MFA’s on the underlying documents.

Algorithm rewrite takes as input an $\mathcal{X}_{\text{reg}}$ query $Q$ and a view definition $\sigma : D \rightarrow D_V$; it returns an MFA $\mathcal{M} = (N_s, \tilde{A}, \tilde{\sigma})$ as output, such that for any XML tree $T$ of $D$, $\mathcal{M}$ on $T$ yields the same result as $Q$ on $\sigma(T)$.

It is based on dynamic programming: for each sub-query $Q'$ of $Q$ and each element type $A$ in $D_V$, it computes a local translation $\text{rewr}(Q', A)$, i.e., an MFA on $D$ that is equivalent to $Q'$ when $Q'$ is evaluated as any $A$ elements of $D_V$. The MFA $\text{rewr}(Q', A)$ is constructed inductively, based on structure of $Q'$. It assembles local translations to obtain $\mathcal{M} = \text{rewr}(Q, r)$, where $r$ is the root type of $D_V$.

Example 5.1: Given query $Q_0$ of Example 4.1 on the view $\sigma_0$ of Example 2.2, assume that we want to compute $\text{rewr}(Q_0, \text{hospital})$. Fig. 5(a) shows a simplified parse tree of $Q_0$. Algorithm rewrite uses this parse tree to inductively build the MFA for $Q_0$. In more detail, Fig. 5(b) shows three MFAs and two AFAs that are the basis of the induction of the rewriting of $Q_0$. Specifically, $\mathcal{M}_0^0$ corresponds to $\text{rewr}(\text{parent}, \text{patient})$, $\mathcal{M}_0^1$ to $\text{rewr}(\text{patient}, \text{parent})$, and $\mathcal{M}_0^2$ to $\text{rewr}(\text{patient}, \text{hospital})$. Notice that the construction of $\mathcal{M}_0^2$ also requires the construction of $A_0^0$.

Figure 5(c) shows how Algorithm rewrite uses these basic blocks to build inductively the MFA $\text{rewr}(Q_0, \text{hospital})$. Specifically, it constructs $\mathcal{M}_0^3 = \text{rewr}(Q_0^0/Q_0^1, \text{hospital})$ by concatenating MFA $\mathcal{M}_0^2$ and $\mathcal{M}_0^3$. Then, it constructs $\mathcal{M}_0^4 = \text{rewr}(Q_0^0/Q_0^1)^*, \text{hospital}$ by concatenating $\mathcal{M}_0^3$ with $\mathcal{M}_0^4 = \text{rewr}(Q_0^0/Q_0^1, \text{parent})$ and adding appropriate $\epsilon$-transitions for the recursion. Finally, the algorithm considers the rewriting of $Q_0^0[\text{q}]$ and concatenates this to MFA $\mathcal{M}_0^4$ to compute the final result.

Similarly rewrite constructs AFA’s for filters $q$, with the following features. (a) For a “path sub-queries” $Q'$ (i.e., of the form $p$ given in Section 2) of $q$, rewrite defines its AFA in same way as MFA for $Q'$. (b) For logical connectives $\land, \lor$, or $\neg$, rewrite connects inductively obtained AFA’s by introducing a new logical state, i.e., an AND, OR, or NOT state. (c) For nested filters, i.e., $q = p[q_1]$ where $q_1 = p'[q_1]$, rewrite constructs a single AFA, rather than nested AFA’s, for $q$, by “concatenating” the AFA’s for $p$ and $q_1$.

Example 5.2: Consider the filter $q_0$ in the query $Q_0$ of Example 4.1. Figure 5(b) shows how its AFA $A^0_q$ is constructed step-wise, by reusing the MFA’s $M^0_q$, $M^1_q$, $M^2_q$ for path sub-queries, and by concatenating these and “local” AFA’s to build $A^0_q$ and $A^1_q$. Note that although $q_0$ contains a nested filter text()='heart disease’, the two filters are combined into a single AFA and no “nested” AFA’s are required.

Concluding, we have the following result, which, in contrast to Corollary 3.3, justifies the use of MFA’s.

Theorem 5.1: Given a view definition $\sigma : D \rightarrow D_V$ and an $\mathcal{X}_{\text{reg}}$ query $Q$ over $D_V$, Algorithm rewrite computes an equivalent MFA of size at most $O(|Q|\sigma|D_V|)$ over the original document in at most $O(|Q|^2\sigma|D_V|^2)$ time.

6 Evaluation Algorithm

To make query rewriting a practical approach it is necessary to be able to efficiently evaluate MFA’s. We next present an evaluation algorithm for MFA’s, referred to as HyPE (Hybrid Pass Evaluation, Fig. 6). Algorithm HyPE takes as input a document tree $T$, a context node $n$ in $T$ and an MFA $\mathcal{M} = (N_s, \tilde{A})$; it outputs $n[\mathcal{M}]$. The desired result $r[\mathcal{M}]$ is obtained by invoking HyPE with the root $r$ of $T$.

A salient feature of HyPE is that it requires only a single top-down pass over the document tree, and a single pass over an auxiliary structure, which in most cases is much smaller than the document tree. It employs several pruning strategies in its top-down pass to avoid visiting irrelevant parts of the tree and the computation of irrelevant AFA’s.

Since any regular XPath query can be transformed into an MFA, HyPE serve as a stand-alone evaluation algorithm for regular XPath, beyond the rewriting context. To the best of our knowledge, HyPE is the first practical algorithm for evaluating regular XPath. Indeed, no practical algorithm has been provided thus far that can be done within a bounded number of tree traversals. For XPath only, a two-pass algorithm was presented in [16]: a bottom-up phase for evaluating filters followed by a top-down phase for selecting nodes. However, it requires a pre-processing step (another scan of the tree) during which the document tree is converted to a special data format (a binary representation of the tree), and the construction of a tree automaton which are more complex than MFA’s and are possibly large. Algorithm HyPE requires neither pre-processing of the data nor the construction of tree automaton. Moreover, in contrast to HyPE, the two-pass XPath evaluation algorithm may have to evaluate filters at nodes in its first phase, although these nodes will not be accessed in its second phase. As will be verified in Section 7, the pruning technique of HyPE speeds up the
In a nutshell, HyPE also constructs an auxiliary DAG structure, called cans (for candidate answers), representing the history of the run of the selecting NFA $N_s$. Vertices in cans will correspond to states in this run for which the associated AFA evaluated to true. Moreover, vertices in cans are possible annotated with a node in $T$ which is potentially in the answer set $n[M]$. A node in $T$ associated with a vertex in cans will be in $n[M]$ if this node is reachable from a node in cans corresponding to an initial state of $N_s$ at context node $n$. This allows for distinguishing between potential and real answer nodes in cans. In the second phase, cans is traversed top-down to identify the real answer nodes. The size of cans is typically much smaller than $T$.

**Example 6.1:** Consider the MFA $M_0$ in Fig. 3 and the tree $T$ shown in Fig. 4. We illustrate how HyPE evaluates $M_0$ on $T$ through the table in Fig. 7. In the figure, we assume that HyPE already traversed, top-down, the left-most patient (node 2) in the tree and we join the execution of HyPE at the point where node 9 is considered (the first row in the table). Each row in the table corresponds to a step in the execution of HyPE during which the node $n$ at the head of the stack $P$ is considered. In the table, we also show (a) $\text{mstates}(n)$, i.e., the $\epsilon$-closures of states in $N_s$ (i.e., the set of states reached by following one or more $\epsilon$ moves), reached by descending to $n$ in $T$; (b) $\text{fstates}(n)$, i.e., a set of states in $A_0^\text{FA}$ such that $\text{fstates}(n) \neq \emptyset$, which is potentially in the answer set $n[M]$ if this node is reachable from a node in cans corresponding to an initial state of $N_s$ at context node $n$. This allows for distinguishing between potential and real answer nodes in cans. In the second phase, cans is traversed top-down to identify the real answer nodes. The size of cans is typically much smaller than $T$.

**Figure 6. Evaluation algorithm for MFA's.**

evaluation of both regular XPath and XPath queries.

In a nutshell, HyPE consists of two phases (not to be confused with two passes of the tree $T$). In the first phase, the tree $T$ is traversed (top-down) depth-first, during which the MFA $M$ prunes away irrelevant subtrees and identifies which AFA’s in $\mathcal{A}$ need to be evaluated at nodes in the tree. Visited nodes are pushed into a stack $P$. This stack is used to evaluate the AFA’s in a synthesized (bottom-up) way. A node is popped from $P$ once all its related AFA’s have been evaluated. The size of $P$ is at most the depth of $T$. During this traversal, HyPE also constructs an auxiliary DAG structure, called cans (for candidate answers), representing the history of the run of the selecting NFA $N_s$. Vertices in cans will correspond to states in this run for which the associated AFA evaluated to true. Moreover, vertices in cans are possible annotated with a node in $T$ which is potentially in the answer set $n[M]$. A node in $T$ associated with a vertex in cans will be in $n[M]$ if this node is reachable from a node in cans corresponding to an initial state of $N_s$ at context node $n$. This allows for distinguishing between potential and real answer nodes in cans. In the second phase, cans is traversed top-down to identify the real answer nodes. The size of cans is typically much smaller than $T$.

**Example 6.1:** Consider the MFA $M_0$ in Fig. 3 and the tree $T$ shown in Fig. 4. We illustrate how HyPE evaluates $M_0$ on $T$ through the table in Fig. 7. In the figure, we assume that HyPE already traversed, top-down, the left-most patient (node 2) in the tree and we join the execution of HyPE at the point where node 9 is considered (the first row in the table). Each row in the table corresponds to a step in the execution of HyPE during which the node $n$ at the head of the stack $P$ is considered. In the table, we also show (a) $\text{mstates}(n)$, i.e., the $\epsilon$-closures of states in $N_s$ (i.e., the set of states reached by following one or more $\epsilon$ moves), reached by descending to $n$ in $T$; (b) $\text{fstates}(n)$, i.e., a set of states in $A_0^\text{FA}$ such that $\text{fstates}(n) \neq \emptyset$, which is potentially in the answer set $n[M]$ if this node is reachable from a node in cans corresponding to an initial state of $N_s$ at context node $n$. This allows for distinguishing between potential and real answer nodes in cans. In the second phase, cans is traversed top-down to identify the real answer nodes. The size of cans is typically much smaller than $T$.

Going back to the figure, the first row of the table indicates two things. First, since $s_9$ is a final state of $N_s$, we know that node 9 is a candidate answer. Second, state $s_9$ is annotated with $A_0^\text{FA}$ and therefore we need to evaluate $A_0^\text{FA}$ to determine whether node 9 is an actual answer. We remember that $A_0^\text{FA}$ needs to be evaluated on node 9 by initializing $\text{fstates}(9)$ with the initial states of $A_0^\text{FA}$. Consider now the second row in the table. Node 10 is in the
top of $\mathcal{P}$. Furthermore, $\text{mstates}(10)$ is obtained by calling function $\text{NextNFAS}tates$ with arguments the $\text{mstates}(9) = \{s_2, s_4\}$ (line 4 in algorithm of Fig. 6). Similarly, $\text{NextNFAS}tates$ computes $\text{fstates}(10) = \{s_{A3}\}$ from $\text{fstates}(9)$ (line 5 in Fig. 6). The fact that $\text{fstates}(10)$ is non-empty tells us that node 10 is relevant for the evaluation of $A_{10}^\mathcal{P}$. The actual evaluation of $A_{10}^\mathcal{P}$ starts when in the head of $\mathcal{P}$ is node 13. At that point, $\text{fstates}(13)$ includes the final state of $A_{10}^\mathcal{P}$ and from that point on $A_{10}^\mathcal{P}$ is evaluated bottom-up. This hybrid mixing of a top-down with a bottom-up evaluation is the distinguishing characteristic of HyPE. Essentially, HyPE uses the former evaluation type to determine when to initiate the latter. When HyPE returns to $\mathcal{P} = \{1, 9\}$ (the dark grey row of the table), the fact that $\text{fstates}(9)$ includes $\{s_{A1} = \text{true}\}$ indicates that the evaluation of $A_{10}^\mathcal{P}$ results in true. Therefore, node 9 is an actual answer. Concerning cans, this is constructed bottom-up. For each node $n$ for which $\text{mstates}(n) \neq \emptyset$, $\text{mstates}(n)$ is connected to the existing cans, each time the subtree below a child of $n$ has been traversed. For example, when $\mathcal{P} = \{1, 9\}$ (dark grey row), $\text{mstates}(9)$ is connected using the transitions in $M_0$ to the cans structure to its left. At this point, notice that by following the path $s_2, s_3, s_4$ we reach node 11 in $T$. Furthermore, through the new state $s_4$ node 9 is also reachable. When the construction of cans is completed (row with dashed box), a traversal of cans starting from the Init nodes shows that nodes 9 and 11 are still reachable and hence are in the answer of $M_0$ on $T$.

**Complexity.** The complexity of HyPE is determined by that of PCans (for constructing cans) and the traversal of cans. PCans needs for each context node $n$ at most $O(|M|)$ time. Moreover, connecting and updating cans takes at most $O(|M|)$ time as well. Hence, the overall time complexity of PCans is $O(|T| |M|)$. Moreover, PCans requires a single scan of the input document $T$ and cans. The space requirement of PCans is dominated by the size of cans, which, although in the worst case is $O(|T| |M|)$, is typically much smaller than $|T|$. Traversing cans takes again $O(|T| |M|)$ time in the worst case. As a consequence:

**Theorem 6.1:** Given an MFA $M$ and tree $T$, HyPE computes $r[M]$ in at most $O(|T| |M|)$ time and space.

Using the evaluation algorithm together with the rewriting algorithm, we obtain a complete practical method for answering queries on (virtual) views. The overall complexity of our method follows from Theorems 5.1 and 6.1.

**Theorem 6.2:** Given an $\mathcal{X}_{\text{reg}}$ query $Q$ on a view of an XML source tree $T$, our query answering method returns the answer to $Q$ in $O(|Q|^2 |\sigma| |D_V|^2 + |Q| |\sigma| |D_V||T|)$ time.

The size $|T|$ of the document is dominant and is typically much larger than the size $|D_V|$ of the view DTD and the size $|\sigma|$ of the view definition $\sigma$; when only $|T|$ is concerned (e.g., if $D_V$ and $\sigma$ are fixed as commonly encountered in practice), our method answers queries in linear-time (data complexity), and in quadratic combined complexity.

**Variants of HyPE.** Although HyPE already performs well in practice (see Section 7), we developed a novel index structure which enables HyPE to skip even more subtrees. In the following, we denote by OptHyPE the version of HyPE which is built on top of the index, and by OptHyPE-C the version of HyPE which uses a compressed version of the index.

**7. Experimental Study**

We have developed a prototype system SMOOQ [10] supporting MFA’s and algorithms rewrite and HyPE (and its variants OptHyPE and OptHyPE-C). In our experiments, we focused on the most time-consuming module of SMOOQ, i.e., the query evaluator. The experiments were conducted on a dual 2.3GHz Apple Xserve with 4GB of memory. For the generation of our datasets, we used ToXGene [1]. We generated XML documents that conform to a recursive hospital database DTD shown in Fig. 1(a), with sizes ranging from 7MB to 70MB, in 7MB increments. Each increment roughly corresponds to adding the medical history of 10,000 patients to our document tree. Therefore, the largest document stores the medical history of approximately 100,000 patients. The maximal depth of the trees is 13. The generated data consist mainly of element nodes, and to a lesser extent of text nodes. Therefore, the size of the document has a direct impact on query evaluation. For example, our smallest document (7MB) consists of 303,714 element nodes vs 151,187 text nodes. The text nodes are used to increase the selectivity of queries but their size is kept to a minimum (so as not to increase the document size).

Using the generated document trees, we conducted two sets of experiments, one regarding XPath evaluation, the
other regarding regular XPath. The reported times are averaged over at least 5 runs of each experiment.

Evaluating XPath Queries. Since regular XPath subsumes XPath, we investigate the performance of HyPE and its variants for the evaluation of XPath queries.

We compared our performance with that of the Java API for XML Processing Reference Implementation (JAXP RI 1.3), which relies on XERces and XALAN [23]. We also compared with JAXP-COMPILE, a version of JAXP that pre-compiles the input query and converts it into a set of Java classes. The two JAXP versions had similar performance and thus we only report one of them.

We ran various types of XPath queries with simple filters on data values, unions of queries, and Boolean combinations of filters. Figure 8 shows the evaluation time for three different types of XPath queries. We show the evaluation time both for queries with result sizes of a few hundreds of nodes (Figures 8(b) and (c)) and queries that return a few thousands of nodes (Fig. 8(a)). For each query type, we report the evaluation time for JAXP, HyPE, OptHyPE and OptHyPE-C. The figures show clearly that our algorithm consistently outperform JAXP by a factor of three for HyPE, and four for OptHyPE and OptHyPE-C. We also observe that in most cases, both optimized versions of HyPE run almost twice as fast as HyPE. Note as well that the performance of OptHyPE-C is almost identical to that of OptHyPE (while OptHyPE-C uses a compressed index).

Evaluating Regular XPath Queries. The second set of experiments investigated the performance of evaluating regular XPath queries with the different versions of HyPE. Existing alternatives rely on a translation of regular XPath into a more powerful query language like XQuery. We conducted a series of experiments following this approach. Specifically, we translated several regular XPath queries into XQuery and evaluated them in GALAX (http://db.bell-labs.com/galax). These experiments consistently showed that the queries in XQuery required considerably more time than their regular XPath counterparts. As a result we omit GALAX from our discussion because even for a simple regular XPath query on the smallest used document tree, GALAX needed more time than HyPE for the same query on the largest tree. Hence, we only focus on the relative performance of our algorithm.

We ran different types of regular XPath queries that involve Kleene star outside a filter, inside a filter, filters inside Kleene stars and combinations thereof. Figure 9 reports the evaluation time for three of these queries. The overall conclusion is consistent with our observations regarding XPath queries. Indeed, OptHyPE and OptHyPE-C show considerable improvement over HyPE.

An interesting observation is that HyPE prunes a substantial number of element nodes. Specifically, HyPE (resp. OptHyPE) prunes, on average, 78.2% (resp. 88%) of the element nodes for our example queries.

8. Related Work

There has been a host of work on rewriting queries posed on XML views to relational queries on top of RDBMS (e.g., [22, 12]). Even in this setting, recursion in the view DTD makes the translation challenging. As observed by [18], most of the existing approaches cannot translate recursive queries over recursive views (two exceptions are [22, 11]).

There has been little work on query rewriting for XML views in the native XML setting where one does not rely on any RDBMS, i.e., the setting considered in this paper. To
our knowledge, the only work in this context is [9], which showed that $\mathcal{X}$ is closed under query rewriting for non-recursive XML views. Our rewriting algorithm given here is the first practical solution to rewriting queries in XPath and its extension regular XPath over recursive XML views.

In [19], regular XPath was introduced and it was shown that a regular XPath query $Q$ can be evaluated over an XML tree $T$ in $O(|Q||T|)$ time, requiring multiple passes over the document tree. A two-pass algorithm for XPath was developed in [16], but it cannot be easily extended to deal with the Kleene star. As observed already in Section 6, even when only XPath is concerned, our evaluation algorithm, HyPE, does not need a pre-processing step (another scan of $T$) that is required by the algorithm of [16], and is more effective in pruning irrelevant nodes when traversing $T$, among other things. To our knowledge, HyPE is the first practical algorithm to evaluate regular XPath queries.

As remarked in Section 1, several automaton formalisms were proposed for processing multiple XPath queries on streaming XML (e.g., [6, 13]). The idea of AFA was explored by [13] for evaluating XPath filters. However, no previous work has attempted to characterize regular XPath in terms of both NFA and AFA in an integrated automaton.

Another line of research concerns view-based query rewriting and answering (see [14] for a survey). Here, given a set of (materialized) views and a query $Q$ on the underlying database, the goal is to answer $Q$ solely on the basis of the views. The problem we consider here is the opposite scenario where the query $Q$ is posed on the view, and it is to find a rewriting $Q'$ of $Q$ on the underlying document.

9. Conclusion

We have provided a solution for efficiently answering regular XPath queries posed on possibly recursively defined XML views. On the theoretical side, we have established results for the closure property and complexity of rewriting (regular) XPath queries on views into (regular) XPath queries on the source. On the practical side, we proposed a practical approach for query rewriting, by using MFA as an intermediate representation of rewritten regular XPath queries. The novelty of the approach consists in (a) an algorithm for rewriting regular XPath queries on XML views to equivalent MFA on the source, and (b) an efficient evaluation algorithm for MFA. These yield an effective method for answering queries posed on XML views of XML data and are useful in enforcing XML security, among other things. Furthermore, our evaluation algorithm is among the first for efficiently processing regular XPath queries. We have fully implemented a prototype system supporting all these algorithms, and our experimental results verified the efficiency of our techniques. We are currently studying extension of our rewriting algorithms to handle XML queries and views defined in XQuery and XSLT.

Acknowledgment. Wenfei Fan is supported in part by EP- SRC GR/S63205/01, GR/T27433/01 and BBSRC BB/D006473/1. Floris Geerts is a postdoctoral researcher of the FWO Vlaanderen.

References