Information Systems 58 (2016) 1–13

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Containment for queries over trees with attribute value comparisons[☆]

Maarten Marx, Evgeny Sherkhonov[*]

University of Amsterdam, Science Park 904, 1098XH Amsterdam, Netherlands

ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form
23 October 2015
Accepted 26 November 2015
Available online 5 December 2015

Keywords:
Tree query languages
Conjunctive queries over trees
Positive XPath
Containment
Attribute

ABSTRACT

Björklund et al. [6] showed that containment for conjunctive queries (CQ) over trees and positive XPath is respectively Π_2^P and coNP-complete. In this article we show that the same problem has the same complexity when we expand these languages with XPath's attribute value comparisons. We show that different restrictions on the domain of attribute values (finite, infinite, dense, discrete) have no impact on the complexity. Making attributes required does have an impact: the problem becomes harder. We also show that containment of tree patterns without the wildcard $*$, which is in PTIME, becomes coNP-hard when adding equality and inequality comparisons.

1. Introduction

In this article we study the containment problem for positive XPath (PosXPath) and conjunctive queries (CQ) interpreted over finite unranked ordered trees with respect to the axes *Child*, *NextSibling*, *Descendant*, NextSibling$^+$ and *Following*. PosXPath is a large fragment of Core XPath [14] which contains all the axes and constructs except negation. Conjunctive queries over trees are an analog of relational conjunctive queries, which correspond to the select-from-where SQL queries in which the where-condition uses only conjunctions of equality comparisons, and are the most widely used query language in practice. A thorough study of the containment problem for CQ over trees has been done in [6]. Their main result is Π_2^P-completeness of the problem. In fact, conjunctive queries can be reformulated as the positive fragment of Core XPath with path intersection. Thus, the Π_2^P hardness result also holds for the containment problem for this fragment. Inspection of the proof in [6] also indicates that the containment for just PosXPath remains in coNP. This extends the result of Miklau and Suciu [19], who showed that containment for tree patterns is coNP-complete.

The query language considered in these previous results ignores attributes. However, in many practical scenarios we deal with data that come from numeric domains, such as real or natural numbers. Thus, it is natural to consider conjunctive queries expanded with attribute value comparisons and study basic static analysis problems such as satisfiability and containment. Such an expansion has been considered for Tree Patterns in [1], where a Π_2^P-completeness result for the containment has been established. However, the hardness proof relies on the construct that allows comparisons of attributes of two different nodes, a feature that is not expressible in Core XPath. As a positive counterpart, a coNP upper bound for containment was shown in the case when comparisons are restricted to either so-called left semi-interval or right

[☆] An extended abstract announcing some of the results of this paper was presented at the 16th International Workshop on the Web and Databases (WebDB), New York, June 23, 2013.

[*] Corresponding to: POSTBUS 94323, 1090 GH, Amsterdam, The Netherlands. Tel.: +31 205258627.

E-mail addresses: maartenmarx@uva.nl (M. Marx), e.sherkhonov@uva.nl (E. Sherkhonov).

http://dx.doi.org/10.1016/j.is.2015.11.003
0306-4379/© 2015 Elsevier Ltd. All rights reserved.

semi-interval attribute constraints. For an attribute a and constant c, an attribute constraint $(@_a \mathrm{op} c)$ is left semi-interval if $\mathrm{op} \in \{<, \leq, =\}$.

This article is an extension of [23], where it was shown that the complexity of containment does not increase for tree patterns expanded with *both* left and right semi-intervals constraints together with inequality constraint. Here we show that essentially the complexity does not change in cases of positive XPath and conjunctive queries over trees. Furthermore, the same upper bounds hold for the cases when we make certain assumptions on the underlying attribute domain D. That is, we show that all the complexity results still hold for the cases when D is a dense or discrete infinite linear order, with or without endpoints, or a finite linear order. As another result, we show that by requiring at least one attribute to be defined in every node of a tree, the complexity of containment over such trees rises to PSpace. If, on the other hand, we require attributes to be defined only at nodes with a certain label (which can be expressed in DTDs) the complexity remains in coNP.

All the upper bound results for both PosXPath and CQ are obtained from a suitable polynomial reduction to the containment problem in PosXPath$^{\neg s}$ and UCQ$^{\neg s}$ (PosXPath and CQ expanded with safe label negation and union) over trees in which nodes may have multiple labels, respectively. Safe label negation is the construct $p \setminus \{q_1, \ldots, q_n\}$ which denotes p-labelled nodes that are not labelled with any of thelabels $q_1, \ldots, q_n$. Table 1 summarizes our results.

The paper is organized as follows. In Section 2 we briefly mention related work. Section 3 contains all the necessary preliminary notions. Section 4 contains the main results. In particular, in Section 4.1 we show that containment for UCQ$^{\neg s}$ and PosXPath$^{\neg s}$ is in Π_2^P and coNP respectively. Next in Section 4.2 we consider containment for CQ$^@$ and PosXPath$^@$ and show the same upper bounds by reducing to the previous problem. Then in Section 4.3 we show that the upper bounds of containment do not change in case of some natural restrictions on the attribute domain. Section 4.4 contains lower bounds: containment of tree patterns without wildcard rises from PTime to coNP when we add equality and inequality comparisons; containment of tree patterns rises from coNP to PSpace when we add equality and inequality comparisons and interpret them on trees in which at least one attribute is defined at each node (a so-called *required* attribute). We finish with conclusions and future work.

Table 1
Complexity results for containment of Positive XPath and CQ with attribute value comparisons.

	PosXPath$^@$	CQ$^@$
No attributes	coNP [6]	Π_2^P [6]
Optional attributes	coNP (Theorem 2)	Π_2^P (Theorem 2)
Required attributes	PSpace-hard (Theorem 3)	PSpace-hard (Theorem 3)

2. Related work

Containment of Conjunctive Queries (CQ) with arithmetic comparisons. The classical result on containment of conjunctive queries over relational databases is its NP-completeness [7]. Later, containment for conjunctive queries expanded with arithmetics comparisons was shown to be Π_2^P-complete [17,26]. In [2], Afrati et al. consider various restrictions on type of comparison operations on either of the two input conjunctive queries with comparisons, as well as on interaction between the comparisons, so that the containment is in NP (cf. Table 1 in [2]). However, it was left open what the exact complexity of containment for CQ with comparisons of type $X \mathrm{op} c$ is, where c is a constant and $\mathrm{op} \in \{=, \neq, \leq, \geq, <, >\}$, i.e., the type of comparisons that we consider in this paper. Note that the Π_2^P-lower bound proof in [26] uses disequation of variables, i.e., the construct $X \neq Y$ for variables X and Y. Nevertheless, adding comparisons of the form $X \mathrm{op} c$ to conjunctive queries does change the complexity of containment, which is in contrast with the result of the current paper for PosXPath and CQ over trees. This is argued in [13], where Π_2^P-hardness of containment for CQ with comparisons was shown, using comparisons of the form $X \neq c$. This proof can also be adapted to use comparisons of the form both $X \leq c$ and $X > c$ [21].

Relational conjunctive queries with negated atoms were also studied previously. It is known that containment for CQ with negated atoms is Π_2^P-complete [25]. The analog of safe negation that we consider here was also considered in the context of relational CQ [27]. In this case, negation in a conjunctive query is *safe* if every variable appearing in a negated atom also occurs in a positive atom of the query. Interestingly, the lower bound proof from [13] can be adapted to show Π_2^P-hardness of CQ with safe negation [21]. Thus, adding safe negation to relational conjunctive queries does change the complexity of containment (from NP to Π_2^P), which is in contrast with the result of the current paper where safe negation does not change the complexity of containment for queries over trees.

Containment for queries interpreted over trees. The containment problem for various XPath fragments has been a topic of wide interest for the past several years. A polynomial time algorithm for tree patterns without the wildcard based on homomorphism between queries was given in [3]. The main result of Miklau and Suciu [19] is the coNP-completeness of containment of tree patterns *with* the wildcard. Almost a complete picture of the containment problem for the XPath fragments with disjunction, in the presence of DTDs and variables was given in [20]. Notably, it was shown that with a finite alphabet the containment problem rises to PSpace. [28] gives decidability results for various fragments with DTDs and a class of integrity constraints. XPath containment in the presence of dependency constraints was studied in [10,11]. All these complexity results are given for forward fragments of XPath. In this paper we consider all the backward axes (parent and ancestor) together with the document order axis (next sibling, following sibling, and following axes and their inverses). Note that in [22] it was shown that every