MODULAR INTERACTIVE XML SCHEMA INFERENCE BASED ON EXTENDED CONTEXT-FREE GRAMMARS

Nela Olšarová
Bachelor Degree Programme (3), FIT BUT
E-mail: xolsar00@stud.fit.vutbr.cz

Supervised by: Zbyněk Křivka
E-mail: krivka@fit.vutbr.cz

ABSTRACT

XML has become a popular format for data exchange and manipulation. The internal structure of XML documents is described by the schema that plays an important role in the data management. The main idea is to develop a method for the schema inference from the given set of XML documents. By observing published automatic methods, we discovered their disadvantages. Since there is no interactive implementation, we improve these methods with the user interaction. As a solid formal background, the principle of grammatical inference is extended in our approach.

1 ÚVOD

XML je v současné době standardem pro uchovávání a přenos informací. Struktura jednotlivých XML dokumentů je popisována schématem, to umožňuje dodržovat jednotnou strukturu dokumentů, které nesou obdobnou informaci. Jazyky XML schémat (DTD, XML Schema a další) jsou ale náročnější na uživatelské ovládnutí než samotné XML, proto se zabýváme návrhem nástroje, který umožní vytvořit k existujícím XML dokumentům odpovídající schéma. Jazykem výsledného schématu bude XML Schema, které upřednostníme kvůli jeho širokým výrazovým schopnostem přesahujícím možnosti jazyka DTD.

2 VYLEPŠENÍ EXISTUJÍCÍCH ŘEŠENÍ

V literatuře nalézáme popisy algoritmů, které odvozují schéma v jazyce DTD nebo XML Schema. Algoritmy odvozující XML Schema bývají sofistikovanější, protože se snaží o co nejlepší využití možností, které tento jazyk nabízí. Nejdále zachází algoritmus, který byl popsán v [3]. Zde prezentovaná metoda z něj z velké části vychází, přidává ovšem interakci s uživatelem, která se dosud neobjevila v žádném prostudovaném algoritmu.

Interakce s uživatelem nám umožní ideálně rozhodnout problém míry zobecnění vytvářeného schématu, kdy se rozhodujeme mezi zachováním detailů vstupních dokumentů a abstrakcí schématu. Pro ilustraci může jít např. o pravidla opakování podelementů. Výsledkem odvozování pomocí stávajících metod je schéma, které je sice z hlediska použitých heuristik a metrik svým

zobecněním optimální, nemusí však odpovídat preferencím uživatele. Hledání optimálního zobecnění založené na metodě odvozování gramatik s interakcí s uživatelem je hlavním cílem vyvíjeného algoritmu.

3 NÁVRH INFERENČNÍHO ALGORITMU

XML schéma popisuje XML dokumenty z hlediska struktury vyskytujících se elementů a jejich podelementů. Další součástí schématu je informace o atributech a jednoduchých datových typech elementů. Základní stromovou strukturu dokumentů, tedy pravidla pro vzájemné vnořování elementů, popíšeme formální gramatikou a aplikujeme metodu odvozování gramatik, odvození pravidel pro atributy a jednoduché datové typy elementů může být řešeno jednoduchou logikou.

3.1 SCHÉMA JAKO ROZŠÍŘENÁ BEZKONTEXTOVÁ GRAMATIKA

Z formálního hlediska pohlížíme na XML jako na bezkontextový jazyk v Chomského hierarchii, který je popsatelný bezkontextovou gramatikou. Popis uspořádání přímých podelementů popisovaného elementu potom může být vyjádřen regulární gramatikou. Získáváme tak rozšířenou bezkontextovou gramatiku. V algoritmu využijeme speciální definici rozšířené bezkontextové gramatiky, která je inspirována [3].

Definice: Rozšířená bezkontextová gramatika (ECFG) je definována pěticí $G = (T, N, D, \sigma, S)$, kde T, N a D jsou disjunktní abecedy terminálů, neterminálů a datových typů. $T = T_1 \cup T_2$ tak, že $T_1 \cap T_2 = \emptyset$ a mezi T_1 a T_2 existuje bijekce. $S \in N$ je počáteční neterminál, σ je konečná množina pravidel ve tvaru $A \rightarrow \alpha$, $A \in N$ a α je regulární výraz složený z *termů*, kde jeden *term* $t \in T_1(N \cup D)T_2$, tento zápis zkracujeme na $T : (N \cup D)$.

Souvislost ECFG a XML Schema včetně příkladů je znázorněna v tabulce 1.

XML Schema	**ECFG**	**Příklad**
název elementu (tag)	terminál	jmeno, kontakt
jednoduchý datový typ	datový typ	`String`
komplexní typ	neterminál	$\langle OsobaTyp \rangle$, $\langle KontaktTyp \rangle$
definice elementu	pravidlo	$\langle OsobaTyp \rangle \rightarrow$ jmeno:`String`
definice komplexního typu	jedno nebo více pravidel	(kontakt:$\langle KontaktTyp \rangle$)*

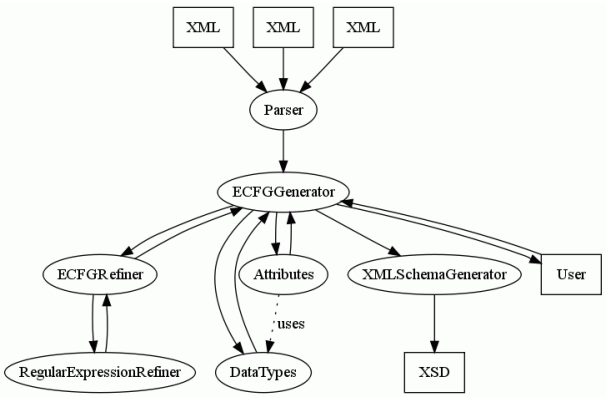
Tabulka 1: Tabulka souvislosti ECFG a XML Schema

3.2 INTERAKTIVNÍ ODVOZOVÁNÍ GRAMATIKY

Výše uvedenou ECFG odvozujeme na základě vstupních dokumentů, které zde slouží jako strukturované pozitivní příklady. V [2] bylo dokázáno, že nekonečný jazyk (kterým je XML) nemůže být odvozen pouze na základě pozitivních příkladů. Interakcí s uživatelem, kterému předkládáme derivační stromy představující lidsky srozumitelnou reprezentaci výsledku, získáváme také příklady negativní. Uživatel tak určuje míru zobecňování. Odvozování regulárního výrazu na pravých stranách pravidel již můžeme provádět i automaticky pomocí slučování stavů konečného automatu, kde aplikujeme poznatky o identifikovatelné podtřídě (k,h)-kontextových jazyků z [1]. I zde je možné zavést obdobnou interakci s uživatelem.

4 MODULÁRNÍ IMPLEMENTACE

Celkové řešení jsme rozdělili do modulů s odpovídající funkčností. Nejprve jsou *Parserem* načteny vstupní dokumenty a je odvozena ECFG, kterou následně při interakci s uživatelem vylepšuje modul *ECFGRefiner* spolupracující s modulem pro odvozování regulárních výrazů na pravých stranách pravidel (*RegularExpressionRefiner*). Součástí odvozování schématu je i upřesnění pravidel pro atributy elementů a odvození jednoduchých datových typů elementů a atributů, toto náleží modulům *Attributes* a *DataTypes*. Konkrétní vhodné metody pro odvozování vlastností atributů a datových typů byly již prezentovány v [3, 4], také zde můžeme využít interakci s uživatelem. V závěru je modulem *XMLSchemaGenerator* vygenerována konečná reprezentace XSD.



Obrázek 1: Schéma nástroje

5 ZÁVĚR

Navrhli jsme novou metodu odvození schématu z XML dokumentů, která má za cíl vylepšit stávající metody o interakci s uživatelem. Při řešení problému využíváme teoretického základu, což přináší možnosti budoucího zkoumání a vylepšování metody i z formálního hlediska.

REFERENCE

[1] Ahonen, H.: Generating Grammars for Structured Documents Using Grammatical Inference Methods. Disertační práce, Department of Computer Science, University of Helsinki, Finland, 1996

[2] Gold, E. M.: Language Identification in the Limit. Information and Control, ročník 10, č. 5, 1967: s. 447-474

[3] Chidlovskii, B.: Schema Extraction from XML Data: A Grammatical Approach. In: KRDB'01 Workshop (Knowledge Representation and Databases), 2000

[4] Hegewald, J., Naumann, F., Weis, M.: XStruct: Efficient Schema Extraction from Multiple and Large XML Documents. In: ICDEW '06, USA: IEEE Computer Society, 2006, ISBN 0-7695-2571-1, 81 s.