Learning restricted regular expressions with interleaving from XML data

1 Definitions for Subclasses

Definition 1. ***SORE*** *[2] Let Σ be a finite alphabet. A single-occurrence regular expression (SORE for short) is a regular expression over Σ in which every terminal symbol occurs at most once.*

Definition 2. ***Simplified CHARE*** *[2] A Simplified CHARE is a SORE over Σ of the form $f_i \cdots f_n$ where $n \geq 1$. Every factor f_i is an expression of the form $(a_1 + \cdots + a_n)$, $(a_1 + \cdots + a_n)^?$, $(a_1 + \cdots + a_n)^+$, $(a_1 + \cdots + a_n)^*$ where $n \geq 1$ and every a_i is a terminal symbol.*

Definition 3. ***eSimplified CHARE*** *[5] An eSimplified CHARE is a SORE over Σ of the form $f_1 \cdots f_n$ where $n \geq 1$. Every factor f_i is an expression of the form $(b_1 + \cdots + b_n)$, $(b_1 + \cdots + b_n)^?$, $(b_1 + \cdots + b_n)^+$, $(b_1 + \cdots + b_n)^*$ where $n \geq 1$ and b_i is the form of a or a^+ where $a \in \Sigma$.*

Definition 4. ***CHARE*** *[1] Base symbols are $a, a^?, a^+, a^+$ where $a \in \Sigma$. A factor is of the form $e, e^?, e^+, e^*$ where e is a disjunction of base symbols of the same kind. A simple regular expression (CHARE for short) is $\emptyset, \varepsilon$, or a concatenation of factors.*

Definition 5. ***eCHARE*** *[6] Base symbols are $s, s^?, s^+, s^+$ where s is a non-empty string. A factor is of the form $e, e^?, e^+, e^*$ where e is a disjunction of base symbols of the same kind. That is of the form $(s_1 + \cdots + s_n)$, $(s_1^? + \cdots + s_n^?)$, $(s_1^+ + \cdots + s_n^+)$, $(s_1^* + \cdots + s_n^*)$, where $n \geq 1$ and s_i is non-empty string. An eCHARE is $\emptyset, \varepsilon$ or a concatenation of factors.*

Definition 6. ***Disjunctive Multiplicity Expression (DME)*** *[3, 4] Let Σ be a finite alphabet. A multiplicity is an element from the set $\{0, 1, ?, +, *\}$. A disjunctive multiplicity expression E is defined as: $E := D_1^{M_1} \& D_2^{M_2} \& \cdots \& D_n^{M_n}$, where $n \geq i \geq 1$ and M_i is a multiplicity. Each D_i is defined as: $D_i := a_1^{M'_1} | a_2^{M'_2} | \cdots | a_k^{M'_k}$, where $k \geq j \geq 1$ and M'_j is a multiplicity and $a_j \in \Sigma$.*

Definition 7. ***SIRE*** *[7] The restricted class of regular expressions with interleaving (RREs) are RE(&) over Σ by the following grammar for any $a \in \Sigma$:*

- $S ::= T \& S | T$
- $T ::= \varepsilon | a | a^* | TT$

The subclass of regular expressions with interleaving (SIREs) are those RREs in which every symbol can occur at most once.

References

1. Bex, G.J., Neven, F., Bussche, J.V.D.: DTDs versus XML Schema: a Practical Study. In: International Workshop on the Web and Databases. pp. 79–84 (2004)
2. Bex, G.J., Neven, F., Schwentick, T., Tuyls, K.: Inference of Concise DTDs from XML Data. In: International Conference on Very Large Data Bases, Seoul, Korea, September. pp. 115–126 (2006)
3. Boneva, I., Ciucanu, R., Staworko, S.: Simple Schemas for Unordered XML. International Workshop on the Web and Databases (2015)
4. Ciucanu, R., Staworko, S.: Learning Schemas for Unordered XML. Computer Science (2013)
5. Feng, X.Q., Zheng, L.X., Chen, H.M.: Inference Algorithm for a Restricted Class of Regular Expressions, vol. 41. Computer Science (2014)
6. Martens, W., Neven, F., Schwentick, T.: Complexity of Decision Problems for XML Schemas and Chain Regular Expressions. Siam Journal on Computing **39**(4), 1486–1530 (2013)
7. Peng, F., Chen, H.: Discovering restricted regular expressions with interleaving. In: Asia-Pacific Web Conference. pp. 104–115. Springer (2015)