Complete Axiom System of Cluster Algebra

Kousuke Fukui[1] and Koji Nakazawa[2]

Graduate School of Informatics, Nagoya University, Japan
{fukuin.gospel, knak}@sqlab.jp

Abstract

Top trees with DAG representation can be used to compress huge tree data such as XML documents. However, one tree can be represented by several top trees, so it is necessary to efficiently decide which top trees represent the same tree for higher compression rate.

In this paper, we give a complete axiom system for the equational theory of top trees, called the cluster algebra. In order to prove the completeness, we introduce a reduction system on cluster algebra, and show the strong normalization and the unique normal form property.

1 Introduction

Tree-structured data such as XML documents are widely used in the world. In many cases, such data become huge, and it is necessary to compress them. DAG is one of the most common compression techniques, in which equal subtrees are shared. However, a lot of real XML data have common parts not as subtrees but as intermediate structures, called *clusters*, and hence they cannot be shared as subtrees in DAGs. For example, in Figure 1 (a), this tree has the common structure $b(b[\,])$, which is not a subtree but forms a cluster.

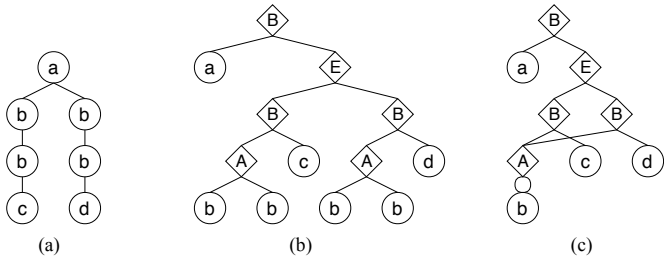

Figure 1: Sharing clusters

To solve this problem, *top trees* and their DAG representations have been proposed [1, 2, 3]. A top tree is a binary tree representing a recipe to reconstructing the original tree by merging its clusters. By the top trees, we can restructure common clusters to subtrees of a top tree, and share them in the top-tree DAGs. In Figure 1 (b), the cluster $b(b[\,])$ is reconstructed as $b\ A\ b$ (we use infix notation for top trees), and it can be shared as a subtree in DAG as (c).

However, the same cluster can be represented by different top trees, and then they cannot be shared in the top-tree DAG. Therefore, if we can efficiently decide whether two top trees represent the same tree, we can expect higher compression rate by the top-tree DAGs.

In this paper, we consider an equational theory for the equivalence of top trees as a theoretical foundation for equivalence checking for top trees. We give an axiom system for the equational theory, called the *cluster algebra* [4], and prove its completeness. For the completeness, we give a reduction system for the cluster algebra, and show the strong normalization and the unique normal form property. We show that there is a one-to-one correspondence between the normal forms and the original trees.

2 Top Tree and Cluster Algebra

We consider ordered trees as original data, in which each of the nodes and leaves has a label, such as $a(b(c,d),e)$. *Clusters* are fragments of a ordered tree. Each cluster has one top boundary node $\bot$ at the root position, and at most one bottom boundary node, which is a distinguished leaf node marked with $[\,]$ such as $a[\,]$. For example, $\bot(b[\,],e)$, $\bot(c,d)$, and $\bot(a(b(c,d),e))$ are clusters in the ordered tree $a(b(c,d),e)$.

We can reconstruct the original ordered tree by merging its clusters. There are five type of merging, which are listed in Figure 2.

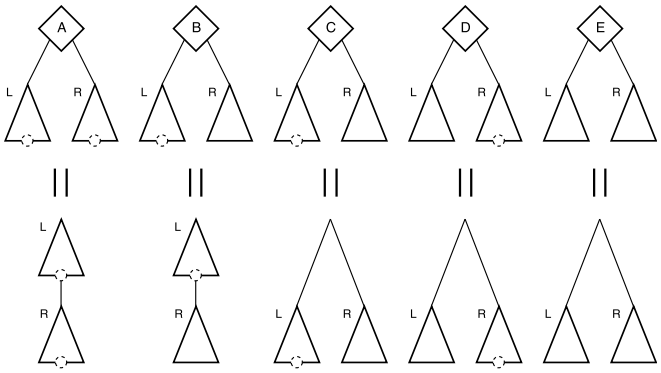

Figure 2: Five types of merging clusters

The type A and B merge two clusters in vertical direction, that is, they replace the bottom boundary node of the left cluster with the right cluster. The difference between A and B is whether the right cluster has a bottom boundary node $[\,]$ or not. The C, D, and E merge two clusters in horizontal direction. For the type C, the left cluster has $[\,]$. For D, the right cluster has $[\,]$. For E, neither has $[\,]$. For example, if two clusters $\bot(b[\,],e)$ and $\bot(c,d)$ are merged by the type B, we obtain the cluster $\bot(b(c,d),e)$. If two clusters $\bot(b,c[\,])$ and $\bot(e)$ are merged by the type C, we obtain

the cluster $\bot(b, c[\,], e)$. We also use the metavariables V for A or B, and H for C, D, or E.

A *top tree* [1, 2] is a binary tree that shows how we can reconstruct the original ordered tree by merging its clusters. Each node of a top tree is one of the merging types A, B, C, D, and E. Each leaf of a top tree is a label of the original ordered tree, which represents either the cluster $\bot(a)$ or $\bot(a[\,])$ depending on the type of the parent node. In [4], the equivalence of the top trees are formalized as the *cluster algebra*, where the merging types are classifying into two kinds, vertical and horizontal merging.

We abuse the metavariables t, $t', \cdots$ for ordered trees and clusters, and we use the metavariables τ, α, $\beta, \cdots$ for top trees. For a cluster t which contains a bottom boundary node and another cluster $t' \equiv \bot(t_1, \cdots, t_n)$, we write $t[t']$ for the cluster obtained by replacing $a[\,]$ in t by $a(t_1, \cdots, t_n)$. For clusters $t = \bot(t_1, \cdots, t_n)$ and $t' = \bot(t'_1, \cdots, t'_m)$, we write $t \odot t'$ for $\bot(t_1, \cdots, t_n, t'_1, \cdots, t'_m)$.

Definition 1. *1. The mapping T from the top trees to the clusters without $[\,]$ and T' from the top trees to the clusters with $[\,]$ are defined as follows.*

$$\begin{aligned} T(a) &= \bot(a) & T'(a) &= \bot(a[\,]) \\ T(\tau_1\ B\ \tau_2) &= T'(\tau_1)[T(\tau_2)] & T'(\tau_1\ A\ \tau_2) &= T'(\tau_1)[T'(\tau_2)] \\ T(\tau_1\ E\ \tau_2) &= T(\tau_1) \odot T(\tau_2) & T'(\tau_1\ C\ \tau_2) &= T'(\tau_1) \odot T(\tau_2) \\ & & T'(\tau_1\ D\ \tau_2) &= T(\tau_1) \odot T'(\tau_2) \end{aligned}$$

The cases which are not listed above are undefined. We call τ well-formed if $T(\tau)$ or $T'(\tau)$ is defined. In the following, we consider only well-formed top trees.

2. When either $T(\tau_1) \equiv T(\tau_2)$ or $T'(\tau_1) \equiv T'(\tau_2)$ holds, τ_1 and τ_2 are said to be equivalent, and we write $\models \tau_1 = \tau_2$.

3 Axioms for Cluster Algebra

We give a set of axioms for equational theory of the cluster algebra.

Definition 2. *1. The axioms for cluster algebra are given as follows.*

$$\begin{aligned} (\alpha\ C\ \beta)\ B\ \gamma &= (\alpha\ B\ \gamma)\ E\ \beta & (\alpha\ E\ \beta)\ E\ \gamma &= \alpha\ E\ (\beta\ E\ \gamma) \\ (\alpha\ C\ \beta)\ A\ \gamma &= (\alpha\ A\ \gamma)\ C\ \beta & (\alpha\ C\ \beta)\ C\ \gamma &= \alpha\ C\ (\beta\ E\ \gamma) \\ (\alpha\ D\ \beta)\ B\ \gamma &= \alpha\ E\ (\beta\ B\ \gamma) & (\alpha\ D\ \beta)\ C\ \gamma &= \alpha\ D\ (\beta\ C\ \gamma) \\ (\alpha\ D\ \beta)\ A\ \gamma &= \alpha\ D\ (\beta\ A\ \gamma) & (\alpha\ E\ \beta)\ D\ \gamma &= \alpha\ D\ (\beta\ D\ \gamma) \\ & & (\alpha\ A\ \beta)\ B\ \gamma &= \alpha\ B\ (\beta\ B\ \gamma) \\ & & (\alpha\ A\ \beta)\ A\ \gamma &= \alpha\ A\ (\beta\ A\ \gamma) \end{aligned}$$

2. We write $\vdash \tau_1 = \tau_2$ if it is derivable by the following inference rules, where $X \in \{A, B, C, D, E\}$.

$$\frac{\tau_1 = \tau_2 \text{ is an axiom}}{\vdash \tau_1 = \tau_2}\ (\textsc{Ax}) \qquad \frac{}{\vdash \tau = \tau}\ (\textsc{Ref}) \qquad \frac{\vdash \tau_1 = \tau_2}{\vdash \tau_2 = \tau_1}\ (\textsc{Sym})$$

$$\frac{\vdash \tau_1 = \tau_2 \quad \vdash \tau_2 = \tau_3}{\vdash \tau_1 = \tau_3}\ (\textsc{Tr}) \qquad \frac{\vdash \tau_1 = \tau_2}{\vdash \tau_1 X \tau = \tau_2 X \tau}\ (\textsc{ComL}) \qquad \frac{\vdash \tau_1 = \tau_2}{\vdash \tau X \tau_1 = \tau X \tau_2}\ (\textsc{ComR})$$

The axioms in the left column exchange V and H. The axioms in the right column are associativity for V and H, respectively.

The soundness is proved by the induction on $\vdash \tau_1 = \tau_2$ straightforwardly.

Theorem 1 (Soundness)**.** *For top trees τ_1 and τ_2, if $\vdash \tau_1 = \tau_2$, then $\models \tau_1 = \tau_2$*

4 Completeness

For the completeness, we introduce a reduction system and prove the strong normalization and the unique normal form property, where we use the fact that there is a one-to-one correspondence between the normal forms and the ordered trees.

Definition 3. *The reduction rules for the cluster algebra are obtained from the axioms in Definition 2.1 by orienting from left to right, such as $(\ \alpha\ C\ \beta\ )\ B\ \gamma\ \Rightarrow\ (\ \alpha\ B\ \gamma\ )\ E\ \beta$*

Theorem 2. *The reduction system for the cluster algebra is strongly normalizable.*

Proof. We define the following three measures.

$$\begin{aligned} w(\tau) &= \Sigma_{V \in \tau}(\text{the number of } H \text{ in the left subtree of } V) \\ d_V(\tau) &= \Sigma_{V \in \tau}(\text{the number of } V \text{ in the left subtree of } V) \\ d_H(\tau) &= \Sigma_{H \in \tau}(\text{the number of } H \text{ in the left subtree of } V) \end{aligned}$$

Then, the pair $(w(\tau), d_V(\tau) + d_H(\tau))$ is strictly decreasing by each reduction step with respect to the lexicographic order. □

The normal forms τ are characterized by the following grammar

$$\tau\ ::= \alpha \mid \alpha\ H\ \tau \qquad\qquad \alpha\ ::= a \mid a\ V\ \tau,$$

Definition 4. *The mapping Θ from the clusters to the normal forms is defined by induction on the size of the clusters as follows.*

$$\begin{aligned} \Theta(\bot(a(t_1, \cdots, t_n))) &= a\ V\ \Theta(\bot(t_1, \cdots, t_n)) \\ \Theta(\bot(t_1, \cdots, t_n)) &= \Theta(\bot(t_1)) H (\cdots (\Theta(\bot(t_{n-1})) H \Theta(\bot(t_n))) \cdots) \end{aligned}$$

Proposition 1. *For any normal form τ, we have $\Theta(T(\tau)) \equiv \tau$.*

The normal forms and the clusters are related as Figure 3.

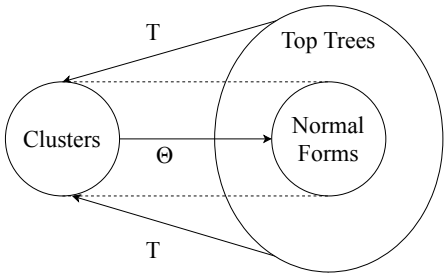

Figure 3: Clusters and normal-form top trees

Proposition 2 (Unique normal form property)**.** *For two normal forms τ_1 and τ_2, if $\models \tau_1 = \tau_2$, then we have $\tau_1 \equiv \tau_2$.*

Proof. By $\models \tau_1 = \tau_2$, we have $T(\tau_1) \equiv T(\tau_2)$, and hence $\Theta(T(\tau_1)) \equiv \Theta(T(\tau_2))$. By the previous proposition, we have $\tau_1 \equiv \tau_2$. □

Theorem 3 (Completeness)**.** *For two top trees τ_1 and τ_2, if $\models \tau_1 = \tau_2$, then $\vdash \tau_1 = \tau_2$.*

Proof. By SN, we have normal forms τ_i' of τ_i for $i = 1, 2$. By the soundness we have $\models \tau_i = \tau_i'$, and by the assumption we have $\models \tau_1' = \tau_2'$. By the previous proposition, we have $\tau_1' \equiv \tau_2'$. Therefore we have $\vdash \tau_1 = \tau_1' \equiv \tau_2' = \tau_2$. □

5 Conclusion

In this paper, we have considered the axioms for the cluster algebra representing the equivalence of the top trees, and proved soundness and completeness of the axiom system. Based on this axiom system, it is expected to be possible to efficiently decide equivalence of top trees without actually decompressing them to the original ordered trees, and higher compression rate in DAG representations of the top trees. As future work, we will give an efficient algorithm for equivalence checker for the top trees.

References

[1] P. Bille, I.L. Gørtz, G.M. Landau, and O. Weimann, Tree compression with top trees, Information and Computation, vol.243, pp.166–177, 2015.

[2] L. Hübschle-Schneider and R. Raman, Tree compression with top trees revisited, Proceedings of the 14th International Symposiumon Experimental Algorithms (SEA2015), LNCS 9125, pp.15–27, 2015.

[3] S. Nishimura, K. Hashimoto, and H. Seki, A Method of Tree Compression with Top Trees and Direct Query Evaluation, IEICE Technical Report 116(127)pp.93–98, 2016. (In Japanese)

[4] A. Gascón, M. Lohrey, S. Maneth, C.P. Reh, and K. Sieber, Grammar-Based Compression of Unranked Trees, Computing Research Repository, `abs/1802.05490`, 2018.