First-Order Theorem Prover Evaluation w.r.t. Relation- and Kleene Algebra

Han-Hing Dang and Peter Höfner

Institut für Informatik, Universität Augsburg, Germany
Han-Hing@gmx.de, hoefner@informatik.uni-augsburg.de

Abstract. Recently it has been shown that off-the-shelf first-order automated theorem provers can successfully verify statements of substantial complexity in relation and Kleene algebras. Until now most of our proof automation had been done using McCune's theorem prover Prover9, while others like SPASS and Vampire were not used extensively. In this paper we use more than 500 theorems to compare and evaluate 13 first-order theorem provers.

1 Introduction

To reach more automation within the areas of relation and Kleene algebras it has been shown that off-the-shelf automated theorem proving (ATP) systems can successfully verify statements of substantial complexity. Examples cover computational logics, relational reasoning, termination analysis, hybrid system analysis and the refinement calculus (see e.g. [6–8]). So far, mainly McCune's Prover9 tool was used for first-order automated reasoning. Motivated by the success and the variety of first-order ATP systems there arises the question whether Prover9 is the best choice for reasoning in such algebras.

Next to this first-order approach, there are others using interactive, higher-order theorem provers (e.g. [9]) or special purpose first-order proof systems [12]. Such approaches require either development effort or user interaction; a disadvantage relative to off-the-shelf first-order ATP systems.

There are general ATP system competitions (e.g. CASC [16]), but obviously these events cannot focus on special algebras. A comparison in the area of demonic refinement algebra with around 40 theorems and 11 theorem provers was done in [8]. It shows that Prover9 and Vampire seem to be the best for that purpose. But, up to now, nobody has done a comparison based on a huge amount of Kleene-like theorems. In total we fed different ATP systems with more than 500 theorems, which allows a reasonable evaluation of the theorem provers involved.

Our main result is that the ATP systems show dramatically different behaviour. On more difficult proof tasks, the best systems are clearly Prover9 and Waldmeister. The latter one can only be used for unit equational logic (i.e., purely equational reasoning). Moreover, we show that the success of theorem provers depends on the encoding of the problems.

In the remainder the expression "*algebra*" will be used as a short hand for "relation and Kleene algebra".

2 The Setting

We have evaluated 13 ATP systems[1] for finding proofs of *algebraic* theorems: Darwin 1.4.1, E 0.999, Equinox 1.3, Geo 2007f, iProver 0.2, leanCoP 2.0, Metis 2.0, Otter 3.3, Prover9 0607, SPASS 3.0, SNARK, Vampire 9.0 and Waldmeister 806[2]. We consistently use a black-box approach to theorem proving, i.e., we do not care about the search strategies or hints the ATP systems use. Initial tasks attempted with all the systems allowed us to select the most powerful systems for our evaluation. In particular, we compared Darwin, E, Otter, Prover9, SPASS, Vampire and Waldmeister in detail. All the other ATP systems failed already in proving some basic properties; it seems that they are not adequate for our task.

For the experiments we used computers with a hyper-threaded 3.0 GHz Intel Pentium 4 CPU and 1 GB memory, running a Linux 2.6 system. We set a CPU time limit of 600 s, which is known to be more than sufficient for the ATP systems to prove almost all the theorems they would be able to prove even with a significantly higher limit [15]. To have a uniform encoding of the involved structures we used the TPTP-format and Sutcliffe's SystemOnTPTP system.

An overview over the results is given in Section 3. Due to lack of space we cannot give the results nor the definitions of the involved structures in full detail. All these as well as the encodings are presented at a website [5].

It is well known that (human) reasoning in variants of Kleene algebras is often inequational. When producing the proof goals, we therefore split equations $s = t$ into inequations ($s \leq t$ and $t \leq s$), but also keep the equational variants, which probably are particularly hard for the ATP systems. Furthermore we also split each equivalence $s \Leftrightarrow t$ into two implications. Finally, implications containing an equation are split in a similar way where possible.

In sum, this yields 500 expressions as proof goals. We have also tested an equational encoding of relational algebra; therefore we used in fact 650 proof goals. We tried to prove them using only the axioms as hypotheses. We do this to have a uniform comparison base for all ATP systems, although it is well known that, in order to obtain ATP proofs of more difficult laws, further lemmas often need to be added to the axioms.

3 The Results

We split the class of all tested goals into three categories: variants of Kleene algebras (e.g. omega algebra or demonic refinement algebra), extensions of Kleene algebras (e.g. modal Kleene algebras) and relation algebras.

Variants of Kleene algebras. This category contains theorems from idempotent semirings, Kleene algebras, omega algebras [2] and demonic refinement algebras [17]. We put these algebras into one single category, since the axiom sets are similar and still small (compared to the next category). Overall we tried

[1] References for the used ATP systems can be found at [5].

[2] Waldmeister is equational based and accepts only universally quantified equations over a many-sorted signature; therefore it cannot directly used for Kleene algebra.

to prove 200 theorems. The proof goals vary from quite simple properties like isotony of addition to very complex formulas like Back's atomicity refinement law [1]. It cannot be expected that any of the ATP systems can prove such a theorem all by itself, since even the shortest proof by hand of this theorem in demonic refinement algebra is almost two pages long and uses lots of auxiliary lemmas. This result was confirmed by our experiments.

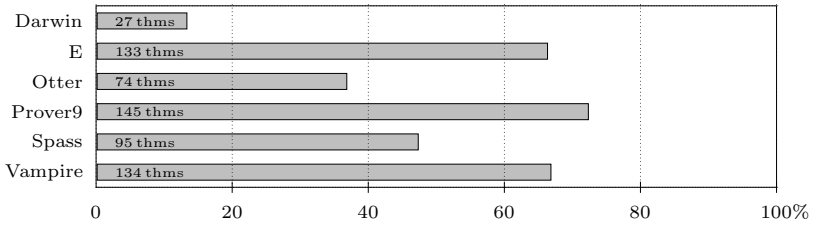

Fig. 1. Results for 200 theorems of variants of Kleene algebra

The results of our experiments are summarised in Fig. 1. The bars illustrate the percentage of theorems proved by the corresponding ATP system. The exact numbers are inside the bars. Although we focused on the most promising ATP systems, these show dramatically different behaviour. The best ones are clearly Prover9, E and Vampire. Overall we think that the success of ATP systems is quite impressive. Prover9 is able to show about 75 % of the goals starting from the axioms only. But, the results confirm that adding lemmas or the use of hypothesis learning techniques are indispensable for ATP systems. In particular, the use of such techniques yields proofs of all more involved theorems (cf. [6–8]).

Analysing the results we recognised that not only human reasoning in *algebra* is often inequational. Also the ATP systems perform much better when splitting equations. Often the systems cannot succeed with equations whereas they are able to prove both inequations. Moreover the different algorithms and strategies of ATP systems behave similar. That means, if Prover9 cannot find a proof, the chance that any other system will succeed is really small.

Extensions of Kleene algebras. This category covers extensions of idempotent semirings, Kleene algebras and omega algebras. In particular, structures with tests [10], with domain [3], and modal structures [14] are included.

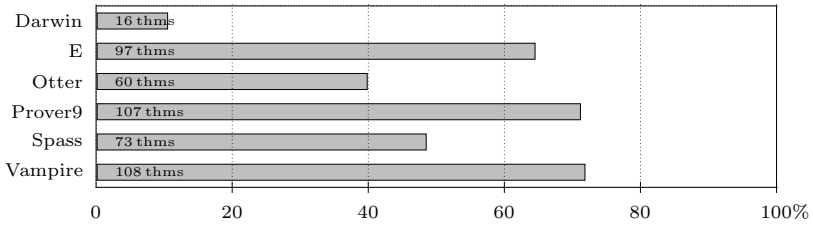

Fig. 2. Results for 150 theorems for extensions of Kleene algebras

The encodings of the extensions are based on the original axiomatisations. For example, modal Kleene algebra is based on the domain operator, which itself is based on a test algebra (a Boolean subalgebra). Such an axiomatisation yields

a dramatical increase in the number of axioms. Therefore, at first sight, these theorems seem to be more complex for ATP systems. But in fact, the results are basically the same as for the basic variants of Kleene algebra.

Relational algebra. Relational algebra can be encoded using universally quantified equations over a single signature. We used this fact to compare an inequational setting with an equational one.

To encode relation algebra we used an axiomatisation of Maddux [13]. In total, we gave 150 theorems to the ATP systems. Whenever an inequality $s \leq t$ arises, we also produced the equivalent form $s + t = t$ as proof goal and skip the order axiom $a \leq b \Leftrightarrow a + b = b$. To eliminate implications we skolemised the hypotheses. This yields an inequational and an equational encoding. The latter was also given to Waldmeister, which is known to be the most powerful system for equational deduction. The results for both encodings are listed in Fig. 3.

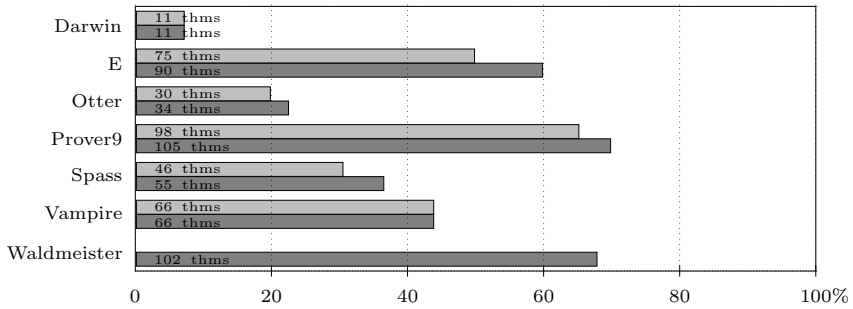

Fig. 3. Results for 150 theorems of relation algebra

The first bar presents the standard encoding, the second the equational one. As a result of our relational experiments we recognised that an equational encoding is significantly better than the standard one. Again the goals cover simple as well as complex theorems. Since we could not transform all theorems into an equational setting, Waldmeister performes even better than illustrated. In fact, it was able to proved 102 out of 135 theorems ($\approx 75\%$). Moreover, the results show that relational algebra is more complex for ATP systems than Kleene algebra. This might be due to the compact axiomatisation of Maddux, or the number of operations directly affects the power of the used systems.

4 Conclusion and Outlook

We tried to find out which theorem provers is the best for our tasks. In total we compared 13 ATP systems and fed them with more than 500 theorems. The best systems are Prover9 and, in the case of an equational encoding, Waldmeister.

But, no ATP system was able to prove all given theorems from the pure axiom set. This is not surprising since we add complex proof goals, like Back's atomicity refinement law. In previous work it has been shown that all the given theorems can be proved with Prover9 [6–8]. The strategies used to reach the goal are various combinations of an increased time limit, adding auxiliary lemmas as additional hypotheses and removal of axioms. For example, adding properties like

transitivity of the ordering, isotony of all operations, or the Dedekind formula would lead to better results. On the other hand, adding all known lemmas would lead to a state space explosion. Therefore axiom selection systems like SRASS become more and more important.

The inclusion of such tools into our experiments is part of our future work. Moreover we want to check whether SPASS performs better if one uses the fact that it can handle relational expressions directly. Furthermore, it will be interesting to analyse the produced results in more detail. For example, we suspect that there exist better *algebraic* encodings. In particular, we want to test a characterisation for the domain operator, which needs no tests [4], and a description for modal Kleene algebra based on modules [11]. Both characterisations yield less axioms and therefore we hope that ATP would yield better results.

Acknowledgements. We thank R. Glück, B. Möller and G. Struth for valuable discussions and comments and G. Sutcliffe for providing the software for TPTP. Last, we are grateful to him and to J. Flierl for their great technical support.

References

1. R.-J. Back. A method for refining atomicity in parallel algorithms. In E. Odijk, M. Rem and J.-C. Syr (eds.), *PARLE '91*, LNCS 366, pp. 199–216. Springer, 1989.
2. E. Cohen. Separation and reduction. In R. Backhouse and J. Oliveira (eds.), *MPC 2000*, LNCS 1837, pp. 45–59. Springer, 2000.
3. J. Desharnais, B. Möller and G. Struth. Kleene algebra with domain. ACM Trans. Comp. Logic, 7(4):798–833, 2006.
4. J. Desharnais and G. Struth. Domain semirings revisited. Technical Report CS-08-01, University of Sheffield, 2008.
5. P. Höfner and G. Struth. <`http://www.dcs.shef.ac.uk/~georg/ka`>.
6. P. Höfner and G. Struth. Automated reasoning in Kleene algebra. In F. Pfennig (ed.), *CADE 2007*, LNAI 4603, pp. 279–294. Springer, 2007.
7. P. Höfner and G. Struth. On automating the calculus of relations. Technical Report CS-08-05, University of Sheffield, 2008.
8. P. Höfner, G. Struth and G. Sutcliffe. Automated verification of refinement laws. (submitted).
9. W. Kahl. Calculational relation-algebraic proofs in Isabelle/Isar. In R. Berghammer, B. Möller and G. Struth (eds.), *RelMiCS 7/AKA 2*, LNCS 3051, pp. 179–190. Springer, 2004.
10. D. Kozen. Kleene algebra with tests. ACM Trans. Program. Lang. Syst., 19(3):427–443, 1997.
11. H. Leiß. Kleene modules and linear languages. J. Log. Algebr. Program., 66(2):185–194, 2006.
12. W. MacCaull and E. Orlowska. Correspondence results for relational proof systems with application to the lambek calculus. Studia Logica, 71(3):389–414, 2002.
13. R. Maddux. Relation-algebraic semantics. Theo. Comp. Sc., 160(1&2):1–85, 1996.
14. B. Möller and G. Struth. Algebras of modal operators and partial correctness. Theo. Comp. Sc., 351(2):221–239, 2006.
15. G. Sutcliffe and C. Suttner. Evaluating general purpose automated theorem proving systems. Artificial Intelligence, 131(1–2):39–54, 2001.
16. G. Sutcliffe and C. Suttner. The state of CASC. AI Comm., 19(1):35–48, 2006.
17. J. von Wright. From Kleene algebra to refinement algebra. In E. A. Boiten and B. Möller (eds.), *MPC 2002*, LNCS 2386, pp. 233–262. Springer, 2002.