First-Order Theorem Provers: the Next Generation
Extended Abstract

Andrei Voronkov
University of Manchester

Abstract

We briefly overview past and present of resolution theorem proving and will try to predict how resolution theorem proving will develop in the next several years. Our presentation will be centered around two main motives: efficiency and usefulness for existing and future potential applications.

From 1993 the author of this paper participates in the development and implementation of the theorem prover Vampire, see [Riazanov & Voronkov 2002a]. The paper is an attempt to look into the future of first-order theorem provers. It is based on the author’s experience with the development, implementation, and applications of Vampire. Since the paper is to be presented at a description logics workshop, we will sometimes refer to reasoning in description logics.

The philosophy of automated reasoning in first-order logic is different from that of reasoning in description logics. First-order logic is undecidable, so first-order theorem proving does not pay much attention to termination. There are decidable classes of first-order formulas but formulas of these classes do not arise often in applications, unless these are classes of formulas obtained by translation from modal or description logics. Due to undecidability, very short formulas may turn out to be extremely complex, while very long ones rather easy. Sometimes first-order provers find proofs consisting of several thousand steps in a few seconds, but sometimes it takes hours to find a ten-step proof. The theory of first-order reasoning is centered around the completeness theorems while in practice completeness is usually not an issue due to the intrinsic complexity of first-order reasoning.

Historically the first description logics turned out to be decidable, and investigation of decidability, complexity, and expressive power remains to be one of the main issues in description logic reasoning. In expressive description logics one has to balance on a narrow margin between decidability and undecidability. When it comes to designing description logic languages, the description logic community tries to stay within the decidable fragments, even when this excludes a class of applications.

Although the two fields seems to stay apart, nevertheless they can learn from one another, especially when it comes to extensions. For example, some connections between the first-order and description logic reasoning were recently considered in [Schmidt 1998, Voronkov 2001, Baader & Tobies 2001].

Introduction. First-order theorem provers have several recognized application areas. These are software and hardware verification, theorem proving in mathematics, first-order knowledge bases, reasoning in nonclassical logics, retrieval of software components, program synthesis etc.
Among the most well-known achievements so far is their use as subsystems of interactive provers, such as HOL, for verification, for example verification of real arithmetic in Intel processors and solution of several open problems mathematics (mainly in algebra), for example the Robbins problem. First-order theorem provers are in practice used much more than it is reported in the literature, since many companies do not disclose proprietary use of the provers.

Since the early work on automated theorem proving, an impressive progress has been made in first-order theorem proving. This progress is due to several factors described below:


3. Growing *experience*, including the experiments with applications and the development of TPTP [Sutcliffe & Suttner 1998].

**Systems.** The main currently supported systems are given in Figure 1. Among these systems, Setheo is based on model elimination, while other provers are based on resolution. In the rest of this paper we will overview resolution theorem provers only.

**Resolution.** In the early papers, e.g., [Robinson 1965, Robinson & Wos 1969] resolution is defined as a logical system consisting of several inference rules operating on clauses, for example, *resolution* and *paramodulation*:

\[
\frac{A \lor C \quad \neg A' \lor D}{(C \land D) \text{mgu}(A, A')} ; \quad \frac{s = t \lor C \quad L[s'] \lor D}{(C \land L[t] \lor D) \text{mgu}(s, s')}.
\]
where mgu denotes the most general unifier. Since these rules are *local*, i.e., their applicability is identified only by a small number of clauses and the result of a rule application does not change the previously generated clauses, resolution is implemented using *saturation algorithms*.

**Saturation algorithms.** These algorithms operate on a set of clauses $S$, initially the set of input clauses. Roughly speaking, they are based on the following loop.

1. Apply inferences to clauses in $S$, adding to $S$ the conclusions of these inferences.
2. If the empty clause $\Box$ is proved, terminate with success. If no inference rule is applicable, terminate with failure.

The set $S$ is the current *search space*. However, a naive application of this loop would not solve even some of the problems considered trivial by the modern theorem provers due to the fast growth of the search space. However, even in the first paper on resolution [Robinson 1965] it was noted that some clauses can be removed from the search space without losing completeness.

**Redundancy.** More precisely, it has been observed that clauses *subsumed* by other clauses can be removed and thus can be regarded as *redundant*. Later, several other notions of redundancy have been discovered. For example, [Brand 1975] proved that the function reflexivity inferences and paramodulation into a variable are redundant. In general, there are two kinds of redundancy in resolution theorem proving: *redundant inferences* and *redundant clauses*. Currently, many useful notions of redundancy are based on *simplification orders* introduced in [Knuth & Bendix 1970]. These orders are orders $\succ$ on terms which can be extended to literals and clauses. An example of an inference which is redundant due to the orderings restrictions is a paramodulation inference

$$s = t \lor C \quad L[s'] \lor D \quad \frac{(C \lor L[t'] \lor D)\theta}{(C \lor L[t'] \lor D)\theta},$$

for which $t\theta \succ s\theta$. In the 1970s–1980s many notions of redundant inferences and clauses have been investigated. In the 1990s a general theory of redundancies was described in [Bachmair & Ganzinger 1994a]. Nearly all state-of-the-art resolution theorem provers are based on this theory. Resolution with redundancy is based on the following saturation algorithm.

1. Apply all *nonredundant* inferences to clauses in $S$, adding to $S$ those conclusions of these inferences that are *nonredundant*.

2. If the empty clause $\Box$ is proved, terminate with success. If no inference rule is applicable, terminate with failure.

3. Remove all sequents that *become redundant* due to the addition of these conclusions of inferences.

In addition to the standard inference rule, these algorithms also operate with *simplifications*. An inference is called a *simplification* if it makes at least one clause in $S$ redundant.
Given clause algorithms. In the modern provers inference selection is mainly done via the clause selection. There are two main concreteizations of the saturation algorithm based on the clause selection: the Otter algorithm [Lusk 1992, McCune 1994] and the Discount algorithm [Denzinger, Kronenburg & Schulz 1997]. These algorithms are described and analyzed in more detail in [Riazanov & Voronkov 2002a].

Search space size and term indexing. Even the proof-search in the inference systems with redundancy creates enormously large such spaces. For example, storing $10^6$ clauses is not unusual. Of these clauses $10^5$ can participate in inferences. Some operations on clauses are very expensive if implemented naively. For example, subsumption on multi-literal clauses is NP-complete and must ideally be performed between each pair of clauses. It is difficult to imagine an implementation able to perform $10^{12}$ operations in reasonable time.

In the state-of-the-art theorem provers all expensive operations are implemented using term indexing [Graf 1996, Sekar et al. 2001]. The problem of term indexing can be formulated abstractly as follows. Given a set $L$ of indexed terms or clauses, a binary relation $R$ over terms or clauses (called the retrieval condition) and a term or clause $t$ (called the query term or clause), identify the subset $M$ of $L$ that consists of the terms or clauses $l$ such that $R(l, t)$ holds.

A typical retrieval condition used in theorem proving is subsumption: retrieve all clauses in $L$ subsumed by $t$. In order to support rapid retrieval of candidate clauses, we need to process the indexed set into a data structure called the index. Modern theorem provers maintain several indexes to support expensive operations. For example, Vampire [Riazanov & Voronkov 2002a] uses flatterms in constant memory for storing temporary clauses, code trees [Voronkov 1995] for forward subsumption, code trees with precompiled ordering constraints for forward simplification by unit equalities, perfectly shared terms for storing clauses, shared terms with renaming lists for backward simplification by unit equalities, path index with compiled database joins for backward subsumption and some other indexes.

Progress in first-order reasoning. Since the first implementation of first-order provers, there has been a considerable progress in solving difficult problems. This progress is due to the development of theory and techniques described above. For example, the provers of the 1970s could not be efficient for equality reasoning because of the unrestricted paramodulation and function reflexivity axioms. The provers of the 1980s would be much faster because of the restrictions on paramodulation and the use of simplification orderings. The provers of the early 1990s paid a special attention to term indexing, and of the late 1990s used the general theory of redundancy, including literal selection functions. Every 10 years, there are problems which are solved by several orders of magnitude faster than by the previous generations of theorem provers.

Efficient theorem proving in future. Theorem proving is a very hard problem. The next generation of theorem provers will incorporate new theory, data structures, algorithms, and implementation techniques to solve the problem which has been characterized by the ANL automatic reasoning group as controlling redundancy in large search spaces [Lusk 1992].

It is unreasonable to expect the future theorem provers to be much faster on all possible problems. However, if we can increase performance of provers by several orders of magnitude for a large number of problems coming from applications, many of these problems will be routinely
solved, thus saving time for application developers. The development of next generation provers will require:

1. development of new theory;
2. addition of new features;
3. development of new algorithms and datastructures;
4. understanding how the theory developed so far can be efficiently implemented on top of existing architectures of theorem provers;

This development is impossible without considerable implementation efforts and a large number of experiments.

The near future. Let me make some predictions about possible directions in theorem proving that will be exploited in the next several years. These predictions are based on the analysis of possible applications of first-order provers. Among the possible applications of theorem provers in the near future are verification, assistance to interactive proof assistants, and semantic web. These applications require development of new features and also increased performance. By improving the existing algorithms and datastructures, we can improve performance on pure first-order problems for some benchmarks by maybe an order of magnitude. With this improved performance many problems will still be beyond the capabilities of provers.

How can we improve their performance by several orders of magnitude? And more importantly, how to improve their performance for practical applications? Several possibilities are sketched below.

Built-in theories. Already in the case of equality it was noted that the naive addition of equality axioms is too inefficient. Development of methods of equality reasoning resulted in an impressive progress of theorem provers. There are other important theories which arise in many applications. Development of specialized reasoning methods for these theories will be a very central problem in theorem proving. This problem is different from the problem of designing decision procedures for theories such as Presburger arithmetic because relations and functions used in these theories may be intertwined with relations and functions from other theories and arbitrary quantifiers. Procedures for the combination of decision and unification algorithms can be of some help [Nelson & Oppen 1979, Shostak 1984, Baader & Schulz 1996, Rueß & Shankar 2001].

Among the most important built-in procedures are the following:

- AC (the theory of associative and commutative functions). These axioms occur in axiomatizations very often. There are many results related to building-in AC in theorem provers but term indexing modulo AC is still in its infancy. Associativity and commutativity were built in the EQP theorem prover [McCune 1997] with a considerable success but without term indexing. Term indexing modulo AC was considered in [Bachmair, Chen & Ramakrishnan 1993] but only for a very special case.

- Theories of transitive relations and orders [Bachmair & Ganzinger 1994b].
• Presburger arithmetic.

As a first step towards building-in important theories one can consider creation of libraries of axioms/theorems about commonly used data types. An example of such a project is the Standard Upper Ontology [IEEE SUO-KIF project 2002]. One can also enhance theorem provers by means of specifying built-in theories, for example, by constraints or by additional inference rules.

Inductively defined types. In many applications of interactive proof assistants one has to deal with inductively defined types. The proof assistants such as Isabelle [Paulson 2002], HOL [Gordon & Melham 1993], COQ [The Coq Development Team 2001], Twelf [Pfenning & Schuermann 1998] have facilities to define data types and functions inductively. First-order theorem provers have no such facilities. The work on building-in inductively defined types can be developed in the following directions:

• Inductive definitions of data types;
• Inductive definitions of functions on these data types;
• Limited forms of inductive reasoning.

Finite domain reasoning. In applications there are often either variables ranging over finite domains or finite relations which can be given by an explicit enumeration of all tuples in them. One can axiomatize membership in a finite domain or a finite relation in first-order logic, but the use of such axioms in a first-order prover would create a large number of axioms. More efficient ways of reasoning with finite domains or relations are required.

Nonstandard quantifiers. A problem similar to the finite domain reasoning is the problem of dealing with quantifiers specifying restrictions on the number of elements satisfying a relation, for example, \( \exists^n \) (there exists at most \( n \)). Such quantifiers are familiar to the description logic community. One can translate formulas with such quantifiers into first-order logic with equality but the translation will create formulas of a prohibitive size. It is interesting to develop special ways of reasoning with such quantifiers, although it is not an easy task.

Working with large axiomatizations. Very often theorem provers must work with large axiomatizations containing many irrelevant axioms. This is typical for applications such as reasoning with ontologies and assisting proof assistants. Recognition of irrelevant axioms is one of the most important problems in theorem proving. Many of the axioms are definitions of relations and functions in various forms. Recognition of the most typical definitions and efficient work with them play a major role in the future provers. The first steps in this direction are reported in [Ganzinger, Nieuwenhuis & Nivela 2001, Afshordel, Hillenbrand & Weidenbach 2001, Degtyarev, Nieuwenhuis & Voronkov 2002].

Proof checking. Surprisingly, the modern theorem provers are not good at producing proofs. Some of them print detailed resolution and paramodulation proofs but none gives a proof of the preprocessing steps, such as skolemization. Without proof-checking one cannot use provers in verification and in assisting proof assistants. We expect that in the near future proof-checking components will be added to all major first-order provers. We conjecture that the
main approaches to proof-checking will be similar to those used in proof-carrying code [Necula & Lee 2000], so that a proof is built using a number of inference rules, and foundational proof-carrying code [Appel 2001], so that a proof will given in a system whose language is rich enough to prove even the preprocessing steps, such as HOL.

Other features. There are other features required of the next generation first-order theorem provers. For example, for the naive users who do not know (and usually do not want to know) much about theorem proving, theorem provers must have a strong auto-mode which will try to select automatically a strategy or strategies best suited for solving the given problem. For more advanced users, there should be options to specify term orderings, literal selection, and clause selection.

References


URL: citeseer.nj.nec.com/denzinger96discount.html


URL: http://www.mpi-sb.mpg.de/SATURATE/Saturate.html


IEEE SUO-KIF project [2002], ‘Standard upper ontology knowledge interchange format’.

URL: http://suo.ieee.org/suo-kif.html


URL: http://www.ai.sri.com/snark/


URL: http://coq.inria.fr/doc/main.html

