Contents

1 Introduction

1.1 Motivation

1.2 Contributions

1.3 Thesis Outline

2 Preliminaries

2.1 First-Order Logic

First-order logic is a formal language used to, amongst others, formalise reasoning, including artifical intelligence, logic programming and automated deduction systems. In this thesis we are only interested in terms. Therefore, we disregard formulas, relations and quantifiers. A more extensive introduction can be found in [1].

A symbol is either a variable, a constant or a function. We choose all variables from the infinite set $\mathcal{V} = \{x, y, z, x_1, x_2, ...\}$, all constants from the infinite set $\mathcal{C} = \{a, b, c, c_1, c_2, ...\}$ and all functions from the infinite set $\mathcal{F} = \{f, g, h, f_1, f_2, ...\}$. Whenever possible, we use only the first three symbols of each set for better readibility.

The arity of a symbol $\text{arity}(s)$ is a positive integer representing the number of arguments the symbol is applied to. All constants have a fixed arity of 0 while every function f has a fixed $\text{arity}(f) \geq 1$. A variable x has an arbitrary but fixed arity depending on its context.

mathrm == no change?

A term in first-order logic, chosen from the infinite set $\mathcal{T} = \{t, u, v, t_1, t_2, ...\}$, is a symbol s applied to $\text{arity}(s)$ arguments, where each argument again is a term. Assume $\text{arity}(f) = 1$ and $\text{arity}(g) = 2$ and, for all other symbols s, $\text{arity}(s) = 0$. Then, the terms $f(a)$ and $g(f(x), a)$ are well-formed while the terms $f(a, b)$, $f(g)$ and $a(b)$ are not.

2.1.1 Generalisation and Unification

Given two terms $t = f(x)$ and $u = f(g(a))$, we are interested in determining whether or not the variable x can be assigned such that $t = u$. If this is indeed possible, we know that t is a more general term than u. In this case, it is trivial to see that $x = g(a)$ implies $t = u$. For the general case, we require some formalised notion of variable assignment.

Definition 2.1. A substitution is a partial function $\rho : \mathcal{V} \longrightarrow \mathcal{T}$. We denote by $t\rho = u$ the term obtained by replacing all variables v in t by $v\rho$ if v is in the domain of ρ. We write $[t_1/x_1, ..., t_n/x_n]$ for the substitution $\{x_1 \mapsto t_1, ..., x_n \mapsto t_n\}$.

When applying the substitution $\rho = [a/x]$ to the term $t = f(x, y)$, we get the term $f(a, y) = t\rho$. As y is not in the domain of ρ, it is not modified in the term. Note that ρ is applied only once to the term, that is, $x[y/x, a/y] = y$ even though y would be substituted by a in the original term.

Definition 2.2. Given two terms t, u, we say that t is a generalisation of u and u an instance or specialization of t if and only if there exists a substitution ρ such that $t\rho = u$. Similarly, we call t and u unifiable if and only if there exists a substitution ρ such that $t\rho = u\rho$. In this case, ρ is called a unifier of t and u.

The question of whether term t is a generalisation of term u is also known as the matching problem in the literature. Similarly, determining whether t and u are unifiable is called the unification problem. [5]

2.1.2 Variable Identity

When solving a matching or unification problem, we must pay attention to variables occuring multiple times. For example, $t = f(x, x)$ is not a generalisation of $u = f(a, b)$ as x cannot be substituted by both a and b. Similarly, $t = x$ is a generalisation of $u = f(x)$ using the substitution $\rho = [f(x)/x]$ but they are not unifiable as $t\rho = f(x) \neq f(f(x)) = u\rho$.

Tracking the identities of the variables while solving a matching problem complicates matters substantially. In this thesis, we will ignore the identity of variables. Instead, we replace them by $*$ and treat each $*$ as a unique variable. For example, the terms $t = f(x, y)$ and $u = f(z, z)$ are both treated as $f(*, *)$ and are therefore not differentiated.

Definition 2.3. Variants are terms identical up to loss of variable identity. In the example, t and u are variants of each other.

Each $*$ is treated as a unique occurence and, while solving a matching problem, we assign each $*$ a unique index. For example, the terms $t = f(x) = f(*_1, a)$ and $u = f(y) = f(b, *_2)$ are unifiable with unifier $\rho = [b/*_1, a/*_2]$.

2.2 Lambda Calculus

The λ-calculus is a formal language used to express computation based on functions. It is defined by a grammar for constructing λ-terms and rules for reducing them. The set of terms $\mathcal{T}$ of the untyped λ-calculus is defined as follows:

1. An infinite set $\mathcal{V}$ of variables. Each variable is a term.

2. If t is a term and x is a variable, then $\lambda x.t$ is a term. This is called an abstraction and represents a function on x.

3. If t and u are terms, then tu is also a term. This is the application of the first argument to the second one.

As we focus mainly on the embedding of first-order terms in λ-terms, the grammar for terms suffices. For a more detailed introduction, see for example [4].

2.3 Isabelle

Isabelle is a generic interactive theorem prover. By design, it uses a metalogic, called Isabelle/Pure, to embed other logics and provide a deduction framework. To do so, Isabelle/Pure uses a higher-order logic. The very basis of this metalogic are simply typed λ-terms within which theorems and inference rules are embedded. [6]

Isabelle is written for the most part in Standard ML (SML) and can also be extended at runtime. It is divided into a small kernel that verifies the correctness of all proofs and the user space within which one can axiomatise new theories and build stronger proof automation.

2.3.1 Term Representation in Isabelle

The λ-terms are a variant of simply typed λ-calculus. They are defined, with minor changes for the sake of simplicity, as follows:

```
datatype term =
    Const of string * typ
  | Free of string * typ
  | Var of string * typ
  | Bound of int
  | Abs of string * typ * term
  | $ of term * term
```

1. `Const` and `Free` both represent a fixed symbol. The latter is used to represent fixed variables in the process of a proof. In this thesis, this distinction is irrelevant: both will be treated as first-order constants.

2. `Var` represents a variable, i.e. it is a placeholder and can be replaced by an arbitrary term of the same type.

3. `Bound` is a variable bound by a lambda term encoded as a de Bruijn index. [2]

4. `Abs` is an abstraction. Although Isabelle uses de Bruijn indices, variables are named for pretty printing purposes.

5. `$` represents the application of the first argument to the second one.

We will ignore the types of terms and simply assume type correctness of all given terms. The λ-term $(\lambda x.\ x)\ a$ can then be represented directly as `Abs x (Bound 1)$ Const a`. The application `$` is written infix and left-associative, i.e. $f\ x\ y$ is written as `Const f $ Var x $ Var y` whereas $f\ (g\ x)$ is written as `Const f $ (Const g $ Var x)`. As there are no tuples in this term representation, all functions are curried by default. That is, `Abs x (Abs y (Const f $ Bound 2 $ Bound 1))` represents the λ-term $(\lambda xy.fxy)$.

We can embed first-order terms in these λ-terms. Variables with an arity of 0 and constants map directly to `Var` and `Const` respectively. Likewise, a function symbol can be represented using `Const`. Terms involving functions are represented by a chain of applications of the constituent subterms. For example, the term $f(a, g(x))$ is represented by `Const f $ Const a $ (Const g $ Var x)`. Note the parentheses around $g(x)$ to differentiate this term from $f(a, g, x)$.

We assume for the sake of simplicity that every term consists of only `Const`, `Free`, `Var` and `$`. Occurrences of `Free` are treated as `Const`. `Abs` are not required for first-order terms and dangling `Bounds`, that is, indices pointing to a non-existing abstraction, are excluded, too.

2.4 Term Indexing

A term index is a datastructure that allows us to efficiently store and query a set of terms. It provides, for example, a *unifiables* query that takes a term index and a term t and retrieves all terms from the term index that are unifiable with t.

Definition 2.4. A term index is an indexed set of terms together with the queries *variants*, *generalisations*, *instances* and *unifiables* that take a term index and a query term and retrieve the corresponding terms from the term index. Terms can be inserted or removed from the indexed set of terms.

A term index groups similar terms in its internal representation. This improves performance for queries by avoiding traversal of each indexed term. For example, when retrieving unifiable terms from the set $\{f(*), f(a), f(g(a)), g(a)\}$ with the query term $g(x)$, the term index can avoid unifiying $g(x)$ with every term with f as top symbol. By failing to unify $g(*)$ with $f(*)$, it will discard all terms with top symbol f and continue directly with $g(a)$.

The term indices differ significantly in their approach to grouping terms. Furthermore, some term indices can also implement other operations efficiently. Some examples, discussed in more depth in [3], are:

1. Merge two indices
2. Retrieve terms unifiable with any term in a query set
3. Return after retrieving the first term

Many specialised operations can be implemented but, alas, we cannot predict which operations will be used. As they can be emulated less efficiently by the simpler operations, we will limit ourselves to the basic query operations, retrieving all the variants, instances, generalisations and unifiables of a query term.

As mentioned in section 2.1.2, we disregard the variable identities. By doing so, we simplify the implementation significantly but obviously obtain incorrect results when retrieving terms. To be precise, the queries will potentially return incorrect terms in addition to the correct terms.

Definition 2.5. A query returning a superset of the correct answer is called an overapproximating query. Similarly, we call a term index overapproximating if it supports only overapproximating queries.

Depending on the context, we may use this overapproximated result either directly or filter the remaining terms another time with a slower, exact method. Handling variable identities correctly in the term index often provides only little benefit and is discussed for a variety of term indices in [3].

3 Term Indexing

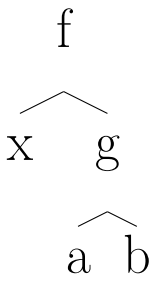
In the following sections we give an overview of path indexing and discrimination trees. We also take a closer look at some details of their implementation in Isabelle/ML as they differ in many places significantly from the approaches chosen in most literature.

3.1 Path Indexing

Instead of storing a term as a tree of functions and their arguments, we can specify the structure and symbols of a tree by combining every symbol of a term with its position in the term, which we call its path. For example, the term $f(x, g(a, b))$ can be represented by the set of paths and their associated symbol that can be seen in fig. 3.1.

Definition 3.1. The path is a sequence of $(symbol, index)$ pairs where the index describes the index of the next argument to traverse.[1]

The paths always start at the top symbol and end with the index at which the symbol is located. For example, $\langle (f, 2), (g, 1) \rangle$ is the path of the symbol a. We represent a path by enclosing a sequence of $(symbol, index)$ pairs with $\langle\rangle$.



- $(\langle\rangle, f)$
- $(\langle f \cdot 1 \rangle, *)$
- $(\langle f \cdot 2 \rangle, g)$
- $(\langle f \cdot 2 \cdot g \cdot 1 \rangle, a)$
- $(\langle f \cdot 2 \cdot g \cdot 2 \rangle, b)$

Figure 3.1: A term and its paths together with the symbols associated with them

Figure: Write as mapping: $\langle\rangle \mapsto f$ etc.

Definition 3.2. $Symbol_t(p)$ refers to the symbol associated with path p in the term t.

Move elsewhere, unnecessary?

[1] This is in contrast to coordinate indexing which only uses a sequence of indices.

A $(path, symbol)$ pair can be interpreted as a constraint on a term where the path defines the position of $symbol$ in the term. For example, $(\langle(f, 1)\rangle, c)$ is only fulfilled by terms of the form $f(c, ...)$. A term gives raise to a set of $(path, symbol)$ pairs, which, when interpreted as constraints, uniquely identify this term up to loss of variable identitification.

These constraints allow us to define terms not explicitly by their structure and symbols but rather by imposing constraints on them. This concept is fundamental to path indexing which stores only the constraints. Queries are resolved by combining the constraints to retrieve a set of terms. For example, a variants query for the term $t = f(x, g(a, b))$ (see fig. 3.1) checks all paths p of t and retrieves all terms u for which $Symbol_t(p) = Symbol_u(p)$.

Move elsewhere

Definition 3.3. A path index is a function $f : Path \times Symbol \longrightarrow 2^{Term}$, that is, each constraint is mapped to a set of terms, which fulfill these constraints and are stored in the path index. We call this set of terms a path set.

Drop "path set" completely?

Storing the path sets such that they can be quickly looked up by a $(path, symbol)$ pair can be achieved in multiple ways. We decided to use a discrimination-tree based approach[2] as many of the paths share prefixes. The nodes of the discrimination-tree contain a function $g : Symbol \longrightarrow 2^{\mathcal{T}}$. The edges are labelled with $(symbol, index)$ pairs, which correspond to the elements of a path.

When we insert a path p of a term t we start at the root and traverse the trie according to p. Once we reach the end of p we extend g of the current node by $Symbol_t(p) \longrightarrow \{t\}$. To insert a term we simply insert all the paths that describe this term. This requires the insertion of many similar paths which profits from the prefix sharing.

If already existant: Expand set, else extend definition of g to $Symbol_t(p)$

Figure 3.2 shows a path index stored as a trie. The root contains a mapping from the symbol f to both terms as they both share this constraint. In the first argument, reached by the edge $(f, 1)$, the symbol a is mapped only to the first term whereas $*$ is mapped to the second term. In the second argument, the terms share the constraint.

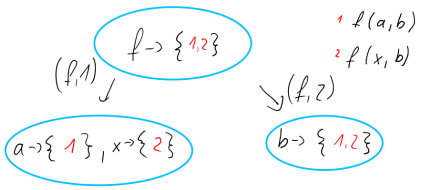


Figure 3.2: A path index

We are interested in retrieving the instances, generalisations and unifiables of a term stored in the index. In addition, we define a lookup to retrieve copies of the term. This can be used to check if a term is already contained but may also be of interest as different variables are not distinguished. The queries are based on intersections and unions of the different path sets to enforce constraints on the terms.

To answer the simplest query, the lookup, we procede as follows:

1. Compute the set of $(path, symbol)$ pairs describing the term.

[2]Not to be confused with the discrimination-tree index which is based on the same datastructure.

2. Retrieve the path sets corresponding to them in the index.

3. Intersect the path sets to retrieve all terms containing the same symbols at identical paths as the query term.

Under the assumption of consistent typing, we retrieve only terms of identical structure as $f(x, y)$ can not exist simultaneously to $f(x)$. Due to the loss of variable identity we may retrieve additional terms.

To retrieve the unifiables of a term from the index, we can use some observations regarding the unification problem.

1. A variable is unifiable with any other term

2. Constants are unifiable with themselves and variables

3. A function $f(x_1, ..., x_n)$ is unifiable with term t if and only if $t = x$ or $t = f(y_1, ..., y_n)$ where for all i x_i is unifiable with y_i. Again, a differing number of arguments are impossible as their types would clash otherwise.

Using this, we can define an algorithm recursing on the structure of the query term while intersecting and unifying the different path sets of the index. The different query types are quite similar, with the lookup being the most restrictive and the unifiables the least restrictive.

PathTerms(p) refers to the path set stored at the path p. AllTerms is the collection of all terms stored in the index.

Arguments \ Query	Lookup	Instances	Generalizations	Unifiables
p x	PathTerms(p·*)	AllTerms	PathTerms(p·*)	AllTerms
p a	PathTerms(p·a)	PathTerms(p·a)	PathTerms(p·*) ∪ PathTerms(p·a)	PathTerms(p·*) ∪ PathTerms(p·a)
p $f(t_1, ..., t_n)$	$\bigcap_n$ Lookup$(p \cdot f \cdot n, t_n)$	$\bigcap_n$ Instances$(p \cdot f \cdot n, t_n)$	PathTerms(p·*) ∪ $\bigcap_n$ Generalisations$(p \cdot f \cdot n, t_n)$	PathTerms(p·*) ∪ $\bigcap_n$ Unifiables$(p \cdot f \cdot n, t_n)$

Figure 3.3: The different queries and their definition

Query Arguments	$Q(p, x)$	$Q(p, a)$	$Q(p, f(t_1, ..., t_n))$
$Q = variants$	$PT(p \cdot *)$	$PT(p \cdot a)$	$\bigcap_i Q(p \cdot f \cdot i, t_i)$
$Q = instances$	$AllTerms$	$PT(p \cdot a)$	$\bigcap_i Q(p \cdot f \cdot i, t_i)$
$Q = generalisations$	$PT(p \cdot *)$	$PT(p \cdot a) \cup PT(p \cdot *)$	$\bigcap_i Q(p \cdot f \cdot i, t_i) \cup PT(p \cdot *)$
$Q = unifiables$	$AllTerms$	$PT(p \cdot a) \cup PT(p \cdot *)$	$\bigcap_i Q(p \cdot f \cdot i, t_i) \cup PT(p \cdot *)$

Figure 3.4: The recursive definition of the queries

3.2 Discrimination Tree

A discrimination tree index, also known as discrimination net index, is a prefix-sharing tree, similar to a trie, which stores terms at its leaves and symbols at its internal nodes. To determine the leaf at which a term is stored we use the preorder traversal of the term. It is obtained by simply reading the written term from left to right. For example, the preorder traversal of $t = f(c, g(x, y))$ is $\langle f, c, g, x, y\rangle$. Since we disregard variable identities, this will further be simplified to $\langle f, c, g, *, *\rangle$.

Definition 3.4. $Preorder(t)$ is the sequence of symbols obtained by the preorder traversal of the term t. For symbols s with $\text{arity}(s) = 0$ it is the symbol s itself. The preorder traversal of a function $f(x_1, ..., x_n)$ is $\langle f, Preorder(x_1), ..., Preorder(x_n)\rangle$ For the sake of simplicity, we flatten the sequence, e.g. $\langle f, \langle g, x\rangle\rangle$ becomes $\langle f, g, x\rangle$.

We store the mapping $Preorder(t) \mapsto t$ in a trie. This way, common prefixes of terms are shared in memory. For example, $Preorder(f(g(x))) = \langle f, g, x\rangle$ shares the first two symbols with $Preorder(f(g(a))) = \langle f, g, a\rangle$. The leafs contain the terms while the internal nodes only store the symbol by which they are addressed. A discrimination tree storing multiple terms can be seen in fig. 3.5. No term is stored in an internal node as, under the assumption of type consistency, it is impossible for $Preorder(t)$ to be a prefix of $Preorder(u)$ if $t \neq u$.

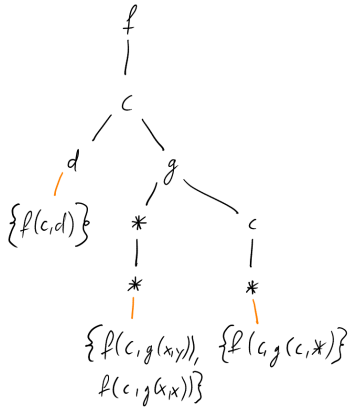


Figure 3.5: Discrimination Tree storing the terms TODO

Root =/= f, besser auf Beispiele ausgelegt: f(c,x) enthalten, netskip kommt vor etc.

The queries are implemented as a recursive algorithm on the nodes of the trie and $Preorder(t)$ of the query term t. Starting at the root, we traverse the tree by selecting the child node corresponding to the first symbol of $Preorder(t)$. We recursively continue at the child node while removing the first symbol from the sequence. In some cases, we may also remove the arguments of the symbol. Therefore, we drop either the first symbol s or $1 + \text{arity}(s)$ symbols from the sequence.

Definition 3.5. $slp(N, c)$ is the symbol lookup operation. It returns the child node of N representing the symbol c. If no such node exists, we return the empty node.

Retrieving the variants of a term is fairly straightforward. Starting at the root, we traverse the trie according to the preorder traversal using *slp*. By doing so, we ensure that we only retrieve terms that contain the same symbols (disregarding variable identity) in

the same order, i.e. the variants of the term.

For example, we retrieve the variants of the term $t = f(c, x)$ in the discrimintation tree fig. 3.5 in the following manner, written as $(Node, Preorder(t))$: $(root, \langle f, c, x\rangle \mapsto (f, \langle c, x\rangle) \mapsto (c, \langle x\rangle) \mapsto (*, \langle\rangle)$ At this point we have reached a leaf containing the term $f(c, x)$, the only stored variant of the term.

The other queries are more intricate as they may now replace variables by arbitrary terms or symbols by arbitrary terms, with unification allowing both. For every constant symbol in the term we form the union of both the query on the node $slp(N, c)$ aswell as $slp(N, *)$. This ensures that indexed terms containing variables are also retrieved.

A variable in the query term must also be handled differently. As the variable may be replaced by arbitrary terms we must skip a number of nodes depending on the arity of the symbol.

Definition 3.6. $skip(N)$ returns the set of nodes obtained by skipping a single term starting at N. For all symbols s for which $slp(N, s)$ is defined, we collect all nodes $1 + \text{arity}(s)$ levels below N. That is, for a constant c with $\text{arity}(c) = 0$ we return $slp(N, c)$ (which is a direct child of N). For a unary function f we return the nodes $slp(slp(N, f), x)$ of all symbols s.

Using this, we can retrieve all the nodes reached by replacing the variable in the query term with some term. The union of the terms returned by the query on each node represents the result. An overview of all the queries is given in fig. 3.6. Note that *variants* is the simplest and most restrictive query, *unifiables* is the most complex and least restrictive with *instances* and *generalisations* being a combination of both.

	$Q(N, \langle\rangle)$	$Q(N, \langle x, t_{2_1} \ldots t_{m}\rangle)$	$Q(N, \langle q, t_{2_1} \ldots t_{m}\rangle)$	$Q(N, \langle f(x_{n_1} \ldots x_{n_n}), t_{2_1} \ldots t_{m}\rangle)$
Q = variants	terms(N)	$Q(slp(N, *), \langle t_{2_1} \ldots t_{m}\rangle)$	$Q(slp(N, q), \langle t_{2_1} \ldots t_{m}\rangle)$	$Q(slp(N, f), \langle x_{n_1} \ldots x_{n_n}, t_{2_1} \ldots t_{m}\rangle)$
Q = instances	terms(N)	$\bigcup_{M \in skip(N)} Q(M, \langle t_{2_1} \ldots t_{m}\rangle)$	$Q(slp(N, q), \langle t_{2_1} \ldots t_{m}\rangle)$	$Q(slp(N, f), \langle x_{n_1} \ldots x_{n_n}, t_{2_1} \ldots t_{m}\rangle)$
Q = generalisations	terms(N)	$Q(slp(N, *), \langle t_{2_1} \ldots t_{m}\rangle)$	$Q(slp(N, *), \langle t_{2_1} \ldots t_{m}\rangle) \cup Q(slp(N, q), \langle t_{2_1} \ldots t_{m}\rangle)$	$Q(slp(N, *), \langle t_{2_1} \ldots t_{m}\rangle) \cup Q(slp(N, f), \langle x_{n_1} \ldots x_{n_n}, t_{2_1} \ldots t_{m}\rangle)$
Q = unifiables	terms(N)	$\bigcup_{M \in skip(N)} Q(M, \langle t_{2_1} \ldots t_{m}\rangle)$	$Q(slp(N, *), \langle t_{2_1} \ldots t_{m}\rangle) \cup Q(slp(N, q), \langle t_{2_1} \ldots t_{m}\rangle)$	$Q(slp(N, *), \langle t_{2_1} \ldots t_{m}\rangle) \cup Q(slp(N, f), \langle x_{n_1} \ldots x_{n_n}, t_{2_1} \ldots t_{m}\rangle)$

Figure 3.6: The definition of the different queries

3.3 Term Indexing in Isabelle/ML

As Isabelle has now been used for over 30 years, a number of data structures have already been implemented to store terms. One of the simplest approaches is the `termtable`, a balanced 2-3 tree, storing terms and differentiating them on all attributes, namely their structure, symbols and types. Therefore, this approach is best used when an exact lookup is necessary. On the other hand, `termtables` do not offer any support for the more complex queries.

The need for efficient retrieval of unifiables and generalisations of a query term is addressed by a discrimination tree implementation. Despite being based on the concept

introduced above, the discrimination tree implementation in Isabelle/ML stores arbitrary sets of values indexed by terms. This allows us to, for example, store horn clauses for efficient backward chaining. By storing the premises of a clause in the node addressed by its conclusion, we can later query our knowledge base for unifiables of our current goal. On successful retrieval, we can replace the goal with the set of premises.

3.3.1 Signature

Before implementing a new term index, we require a unified interface that abstracts the details of the already existing discrimination tree implementation. The result can, with minor omissions, be seen in section 3.3.1. We see that the term index, `T`, is parameterized over the generic and potentially uncomparable `'a`. There are two functions modifying the index. `insert` works as one would expect. The first argument

```
type 'a T
val empty: 'a T

val insert: ('a * 'a -> bool) -> term * 'a -> 'a T -> 'a T
val delete: ('a -> bool) -> term -> 'a T -> 'a T

val content: 'a T -> 'a list
val variants: 'a T -> term -> 'a list
val generalisations: 'a T -> term -> 'a list
val instances: 'a T -> term -> 'a list
val unifiables: 'a T -> term -> 'a list
```

Figure 3.7: The term index signature

As we can see from the signature in section 3.3.1, the discrimination tree stores a list of arbitrary values in each leaf node. The internal nodes, named `Net`, store the children separated into three separate trees. `var` stores both variables and abstractions and atoms stores the constant symbols. Due to the applicative nature of the terms, we also need a

continue here

The generalisation of storing terms to storing arbitrary single values is relatively simple for discrimination trees but, unfortunately, significantly more difficult for path indexing. This compounds with the fact, that the discrimination tree implementation does not only store arbitrary values but instead efficiently stores sets of arbitrary values. While this does not complicate the datastructre (after all, a data structure for arbitrary values can also store sets of values), the semantics of `insert`, `delete` and duplicate detection are consequently more complicated.

We illustrate this with some examples. We use $(term, value)$ for the items stored and DT for the (initially) empty discrimination tree. We use the syntactic equality although we can avoid deduplication by using a non-reflexive comparsion function:

1. Inserting $(a, true)$ and $(b, true)$ into DT stores $true$ at both a and b. Retrieving the unifiables of x returns the multiset $true, true$ as both a and b are unifiable and the queries do not deduplicate the results.

2. Inserting $(x, true)$ and $(y, true)$ into DT results in an exception as both are stored in the same node of the tree and the values are identical.

3. Inserting (x, x) and (y, y) into DT stores both variables x and y in the same node as, by α-equivalence, the values are different.[3]

4. After inserting $(x, true)$ into DT, we cannot delete this value without knowing the term used to address the node where the value is stored. Unfortunately, we can delete this value not only with the term x but also with y and $\lambda x.x$ as these are considered to be equivalent by the index. Additionally, deleting a value completely, that is, from all locations it is stored at, is not possible.

The above may not seem too surprising and can be seen as a natural consequence of the design constraints. For example, storing sets of values nicely sidesteps the problem of distinguishing higher-order terms. The discrimination tree can handle higher-order terms without any problems by mapping all λ-expressions to variables. Therefore, it will return, perhaps extreme, overapproximations to queries but will not fail to store the values or erroneously detect duplication.

Mention mapping of HOL to FOL where?

The lack of deduplication in queries is a consequence of ease-of-use and consistency. As the insertion of the identical value at different nodes succeeds the different instances of the value should be treated differently. Furthermore, storing the same value multiple times is most likely a rare occurence in which the user is responsible for correct handling.

Nevertheless, this represents significant complexity which, while natural for discrimination trees, must be carefully reproduced in the path index implementation. We recall that a term is never explicitly stored in path indexing as we represent a term by a collection of paths which each store a reference to the term. Naively replacing the reference to the term by a reference to a set of values changes the semantics significantly. Different terms often share at least one path but by no longer storing a reference to the term we lose the information from which term a $(path, value)$ pair originates. Implementing the deletion correctly is no longer possible.

Continue here

Unfortunately most literature on path indexing only covers the storage of terms. The queries of path indexing rely on the intersection and union of the path sets. These in turn rely on the fast comparison of the stored values. For example, storing hash tables in the path index would be extremely slow as they cannot be compared directly[4] and comparing the contents requires the collection of all entries. To solve this potential problem we investigated multiple solutions.

2 Quellen sicher, noch welche?

Furthermore, the insertion of an identical $(term, value)$ pair raises an exception. We require a comparison function during insertion to determine this.[5] As the index only stores the terms in the path sets we have to store $(term, value)$ pairs in the path sets. Assume we store only the values in the path sets. We cannot determine whether, for example, the path index containing $(f(c), 1)$ and $(g(d), 1)$ has also stored the $(f(d), 1)$ as the value 1 is present in all path sets associated with $f(d)$, namely $(\langle\rangle, f)$ and $(\langle f, 1\rangle, d)$.

[3] The term equality used by the discrimination tree to map the higher-order terms to the internal first-order terms without variable identity is not exposed.

[4] The structure of the hashtable depends on the insertion order

[5] The comparison function need not be reflexive, for example the constant (λx.false) is valid.

The first approach requires a comparison function for the values.[6] Using the index becomes more difficult by doing so. A user has to implement a comparison function for values and additionally has to consider the potential performance impact. This can be partially mitigated by using pointerEq, although it can only be used as a shortcut for identical values. The comparison must still be called for differing values since there is no perfect sharing.

Code references

Remove paragraph? Doesn't work with (term,value) pairs

The second approach is the storage of $(term, value)$ pairs. By doing so we can implement all the operations according to the literature and simply discard the term before returning the results. This simplifies implementation and retains acceptable performance as the comparison of differing terms will likely only need to compare the first few symbols. It will also increase the memory consumption as a copy of every term is stored solely for the set operations. (Additionally there is no immutable pointer implementation in Isabelle/ML. Instead, copies of identical values are shared by the runtime.)

PolyML erwähnen?

Überhaupt nicht erwähnen weil Data Sharing gut genug funktioniert?

This approach can be further optimised by replacing the $(term, value)$ pairs by $(identifier, value)$ pairs and mapping each term to an identifier. By using integers as identifier, we reduce the comparison to an integer comparison. Additionally, we can use ordered lists, provided by the SML standard library, for the path sets to implement the set operations more efficiently. We are also less reliant on the pointer equality provided by Poly/ML and runtime details like the merging of identical immutable values. This is quite important as we do not have any guarantee when the last heap compression occurred and manual invocation by using the shareCommonData introduces significant overhead to insertion. Additionally, reliance on low-level functions like shareCommonData and pointerEq should be avoided as there are may be significant changes across runtime versions.

Reference

We can further speed up the set operations by building a tree of the intersections and unions and only evaluating it at the end. This likely utilizes the cache better because the previously calculated list is not evicted from the cache by the trie traversal. Furthermore, this presumably enables further compiler optimizations as the intermediate results are only short-lived and functions can be inlined.

Example, too vague

Data Sharing in Poly/ML. NoConstraint exc. No generic hash. Saving "Copy" of values because pointers/ref are always mutable and bad for GC etc.

3.3.2 Delete Semantics

There already exists an implementation of a term index in Isabelle/ML with a two caveats. Firstly,

[6] Either as an argument to every function or by implementing path indexing as a functor on a value module

4 Testing in Isabelle/ML

4.1 Previous Work

A testing framework was built which tries to mirror QuickCheck from Haskell. Lack of typeclasses => Either use functors (OCaml, verbose), Compiler directly (Write tests as strings, no editor assistance, somewhat awkward) or explicitly pass generators, shows, shrinks etc.

Last option best because: Default generators can be provided but more complicated tests need custom generators anyway to satisfy preconditions etc., simple to use, simple to extend.

Modularity was relatively bad

4.2 Term Generation

Approaches: Random or deterministic Carry state around? Yes as we want maximum control. Possibly bad for performance as no multithreading this way. Use symbol generator to give maximum control over structure. Discarded idea: Separate generation into structure and symbols. Too intertwined to be efficiently separated. Address symbols by: Level + Index in Level or Path from root to symbol Deterministic generator with non-deterministic symbol generator works best => Useful for all cases but also simple to use. Symbol generator contains real complexity, term gen relatively basic (basically fold over yet to be generated tree)

4.3 Implementation Details

4.3.1 Overview of Modules

Shrinking: Generate simpler test cases from failed tests. Simplify repeatedly until no longer possible. Depth-First with only one level of Backtracking (or rather none and one consistency check before descending into child). Performance sensitive. Unfortunately the shrinking function must be provided. General shrinking is difficult as generator take no size argument. Else, we could simply take each involved generator and shrink their size (i.e. shrinkg listgen (itemgen 3) 10 returns [listgen (itemgen 2) 10, listgen (itemgen 3) 9]). Combinatorial explosion so not possible. Potential solution: Use compiler here as this is not exposed to the user if no shrink is given (by providing default shrinks for each type and applying them to the provided gen. Would require transparent generators where we can determine in retrospect which symbol gen was used for termgen etc.)

Output Style: Multiple output styles possible. No textfile output at the moment

Inputs: Lazy for performance, can take pregenerated lists of generators e.g. read from file. Constant values are also possible (e.g. ensure that previous failures do not fail again <- Not easily possible but with custom output should be doable: Append failed tests to file, rerun them for every test)

Lehman(?)-PRNG implemented, works as expected.

4.4 Usage

Generators take states. This state is often only a random variable but may contain other values. Generators can take other generators as argument and pass the state around. Deterministic gens dont require random values. Examples

4.5 Usage in this Thesis

What tests were written? What generators used?

5 Evaluation

5.1 Testweise

5.2 Vergleich von Queries, mit DN, Itemnets

5.3 Vergleich von Insert, Delete, (Merge); mit DN, Itemnets, Termtabs

5.4 Wann PI hernehmen statt dem Rest?

5.5 Shortcomings

5.6 Future Work

6 Conclusion

Bibliography

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of databases*. Reading, Mass: Addison-Wesley, 1995. 685 pp. ISBN: 978-0-201-53771-0.

[2] N. G. D. Bruijn. "LAMBDA CALCULUS NOTATION WITH NAMELESS DUMMIES, A TOOL FOR AUTOMATIC FORMULA MANIPULATION, WITH APPLICATION TO THE CHURCH-ROSSER THEOREM". In: (), p. 12.

[3] D. Knuth. "Comparison of indexing techniques". In: *Term Indexing*. Ed. by P. Graf. Red. by J. G. Carbonell, J. Siekmann, G. Goos, J. Hartmanis, and J. Leeuwen. Vol. 1053. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 201–231. ISBN: 978-3-540-61040-3 978-3-540-49873-5. DOI: 10.1007/3-540-61040-5_16.

[4] R. Loader. "Notes on Simply Typed Lambda Calculus". In: (), p. 39.

[5] W. McCune. "Experiments with discrimination-tree indexing and path indexing for term retrieval". In: *Journal of Automated Reasoning* 9.2 (Oct. 1992), pp. 147–167. ISSN: 0168-7433, 1573-0670. DOI: 10.1007/BF00245458.

[6] M. Wenzel. *The Isabelle/Isar Reference Manual*. Feb. 20, 2021.