Techniques for understanding legacy software systems
Kuipers, T.

Citation for published version (APA):
Kuipers, T. (2002). Techniques for understanding legacy software systems
Chapter 5

Types and Concept Analysis for Legacy Systems

In the previous chapter a starting point for an object-oriented re-design was obtained by looking at the different data elements used in a system. In this chapter, this notion is refined. Not the individual data elements are used to perform concept analysis on, but rather the inferred types of the data elements. Type inference tries to identify a group of data elements which are of the same type, i.e., represent the same notion in a system. Using type inference, a large group of data elements can be reduced to a smaller group of data types.

As this chapter has been published as a separate article earlier, we ask the reader to bear with us as some paragraphs in this chapter overlap with paragraphs in the previous chapter. ¹

5.1 Introduction

Most legacy systems were developed using programming paradigms and languages that lack adequate means for modularization. Consequently, there is little explicit structure for a software engineer to hold on to. This makes effective maintenance or extension of such a system a strenuous task. Furthermore, according to the Laws of Program Evolution Dynamics, the structure of a system will decrease by maintenance, unless special care is taken to prevent this [BL76].

Object orientation is advocated as a way to enhance a system’s correctness,

robustness, extendibility, and reusability, the key factors affecting software quality [Mey97]. Many organizations consider migration to object oriented platforms in order to tackle maintenance problems. However, such migrations are hindered themselves by the lack of modularization in the legacy code.

A software engineer's job can be relieved by tools that support remodularization of legacy systems, for example by making implicit structure explicitly available. Recovering this information is also a necessary first step in the migration of legacy systems to object orientation: identification of candidate objects in a given legacy system.

The use of concept analysis has been proposed as a technique for deriving (and assessing) the modular structure of legacy software [DK99b, LS97, SR97]. This is done by deriving a concept lattice from the code based on data usage by procedures or programs. The structure of this lattice reveals a modularization that is (implicitly) available in the code.

For many legacy applications written in Cobol, the data stored and processed represent the core of the system. For that reason, many approaches that support identification of objects in legacy code take the data structures (variables and records) as starting point for candidate classes [CDDF99, FRS94, NK95]. Unfortunately, legacy data structures tend to grow over time, and may contain many unrelated fields at the time of migration. Furthermore, in the case of Cobol, there is an additional disadvantage: since Cobol does not allow type definitions, there no way to recognize, or treat, groups of variables that fulfill a similar role. We can, however, infer types for Cobol automatically, based on an analysis of the use of variables [DM98]. This results in types for variables, program parameters, database records, literal values, and so on, which are used during analysis.

In this paper, we use the derived type information about the legacy system as input to the concept analysis. This way, the analysis is more precise than when we use variables or records as inputs. The concept analysis is used to find candidate classes in the legacy system. External knowledge of the system can be used to influence the concepts that are calculated through ConceptRefinery, a tool we have implemented for this purpose.

All example analyses described are performed on Mortgage, a relation administration subsystem of a large mortgage software system currently in production at various banks. It is a 100.000 LOC Cobol system and uses VSAM files for storing data. The Mortgage system is described in more detail in [DK98, DM99].

5.2 Type inference for COBOL

Cobol programs consist of a procedure division, containing the executable statements, and a data division, containing declarations for all variables used. An example containing typical variable declarations is given in Figure 5.1. Line 6 contains a declaration of variable STREET. Its physical layout is described as picture X(18),
which means "a sequence of 18 characters" (characters are indicated by picture code X). Line 18 declares the numerical variable N100 with picture 9(3), which is a sequence of three digits (picture code 9).

The variable PERSON in line 3 is a record variable. Its record structure is indicated by level numbers: the full variable has level 01, and the subfields INITIALS, NAME, and STREET, are at level 03. Line 12 declares the array A00-P0S: it is a single character (picture X(01)) occurring 40 times, i.e., an array of length 40.

When we want to reason about types of variables, Cobol variable declarations suffer from a number of problems. First of all, it is not possible to separate type definitions from variable declarations. As a result, whenever two variables have the same record structure, the complete record construction needs to be repeated.\textsuperscript{2} Such practices do not only increase the chance of inconsistencies, they also make it harder to understand the program, since a maintainer has to check and compare all record fields in order to decide that two records indeed have the same structure.

In addition, the absence of type definitions makes it difficult to group variables that are intended to represent the same kind of entities. On the one hand, all such variables will share the same physical representation. on the other hand, the converse does not hold: One cannot conclude that whenever two variables share the same byte representation, they must represent the same kind of entity.

Besides these problems with type definitions, Cobol only has limited means to indicate the allowed set of values for a variable (i.e., there are no ranges or enumeration types). Moreover, Cobol uses sections or paragraphs to represent procedures. Neither sections nor paragraphs can have formal parameters, forcing the programmer to use global variables to simulate parameter passing.

To remedy these problems, we have proposed to infer types for Cobol automatically, by analyzing their use in the procedure division. In the remainder of this section, we summarize the essentials of Cobol type inferencing: a more complete presentation is given in [DM98]. First, we describe the primitive types that are distinguished. This is followed by a description of the type relations that can be derived from the statements in a single Cobol program, and how this approach can be extended to system-level analysis leading to inter-program dependencies. Finally, we show how the analysis can be extended to include types for literals, discuss the notion of pollution, and conclude with an example.

**Primitive Types** The following three primitive types are distinguished: (1) elementary types such as numeric values or strings; (2) arrays; and (3) records. Every declared variable gets assigned a unique primitive type. Since variable names qualified with their complete record name must be unique in a Cobol program, these names can be used as labels within a type to ensure uniqueness. We qualify variable names with program or copybook names to obtain uniqueness at the system

\textsuperscript{2}In principle the COPY mechanism of Cobol for file inclusion can be used to avoid code duplication here, but in practice there are many cases in which this is not done.
DATA DIVISION.
/ variables containing business data.
01 PERSON.
   03 INITIALS PIC X(05).
   03 NAME     PIC X(27).
   03 STREET   PIC X(18).
...
/ variables containing char array of length 40,
/ as well as several counters.
01 TAB000.
   03 AOO-NAME-PART.
   05 AOO-POS   PIC X(01) OCCURS 40.
   03 AOO-MAX   PIC S9(03) COMP-3 VALUE 40.
   03 AOO-FILLED PIC S9(03) COMP-3 VALUE 0.
...
/ other counters declared elsewhere.
01 N000.
   03 N100     PIC S9(03) COMP-3 VALUE 0.
   03 N200     PIC S9(03) COMP-3 VALUE 0.

PROCEDURE DIVISION.
/ procedure dealing with initials.
   820-INITIAL SECTION.
   MOVE INITIALS TO AOO-NAME-PART.
   PERFORM R300-COMPOSE-NAME.

/ procedure dealing with last names.
   8230-NAME SECTION.
   MOVE NAME TO AOO-NAME-PART.
   PERFORM R300-COMPOSE-NAME.

/ procedure for computing a result based
/ on the value of the AOO-NAME-PART.
/ Uses AOO-FILLED, AOO-MAX, and N100
/ for array indexing.
   R300-COMPOSE-NAME SECTION.
   PERFORM UNTIL N100 > AOO-MAX
   ...
   IF AOO-FILLED = N100
   ...

Figure 5.1: Excerpt from one of the Cobol programs analyzed (with some explanatory comments added).
level. In the remainder we will use $T_A$ to denote the primitive type of variable $A$.

**Type Equivalence** From expressions that occur in statements, an equivalence relation between primitive types is inferred. We consider three cases: (1) relational expressions: such as $v = u$ or $v \leq u$, result in an equivalence between $T_v$ and $T_u$; (2) arithmetic expressions: such as $v + u$ or $v * u$, result in an equivalence between $T_v$ and $T_u$; (3) array accesses: two different accesses to the same array, such as $a[v]$ and $a[u]$, result in an equivalence between $T_v$ and $T_u$.

When we speak of a type, we will generally mean an equivalence class of primitive types. For presentation purposes, we will also give names to types based on the names of the variables part of the type. For example, the type of a variable with the name L100-DESCRIPTION will be called DESCRIPTION-type.

**Subtyping** From assignment statements, a subtype relation between primitive types is inferred. Note that the notion of assignment statements corresponds to Cobol statements such as MOVE, COMPUTE, MULTIPLY, etc. From an assignment of the form $v := u$ we infer that $T_u$ is a subtype of $T_v$, i.e., $v$ can hold at least all the values $u$ can hold.

**System-Level Analysis** In addition to type relations that are inferred within individual programs, we also infer type relations at the system-wide level: (1) Types of the actual parameters of a program call (listed in the Cobol USING clause) are subtypes of the formal parameters (listed in the Cobol LINKAGE section). (2) Variables read from or written to the same file or table have equivalent types.

To ensure that a variable that is declared in a copybook gets the same type in all programs that include that copybook, we derive relations that denote the origins of primitive types and the import relation between programs and copybooks. These relations are then used to link types via copybooks. ³

**Literals** An extension of our type inference algorithm involves the analysis of literals that occur in a Cobol program. When a literal value $l$ is assigned to a variable $v$, we infer that the value $l$ must be a permitted value for the type of $v$. Likewise, when $v$ and $l$ are compared, value $l$ is considered to be a permitted value for the type of $v$.Literal analysis infers for each type, a list of values that is permitted for that type. Moreover, if additional analysis indicates that variables in this type are only assigned values from this set of literals, we can infer that the type in question is an enumeration type.

³Another (possibly more precise) approach would be to derive a common supertype for all versions that appear in different programs. Our case studies, however, showed no need for such an approach.
Aggregate Structure Identification When the types of two records are related to each other, types for the fields of those records should be propagated as well. In our first proposal [DM98], we adopted a rule called substructure completion, which infers such type relations for record fields whenever the two records are identical (having the same number of fields, each of the same size). Since then, both Eidorff et al. [EHM+99] and Ramalingam et al. [RFT99] have published an algorithm to split aggregate structures in smaller “atoms”, such that types can be propagated through record fields even if the records do not have the same structure.

Pollution The intuition behind type equivalence is that if the programmer would have used a typed language, he or she would have chosen to give a single type to two different Cobol variables whose types are inferred to be equivalent. We speak of type pollution if an equivalence is inferred which is in conflict with this intuition.

Typical situations in which pollution occurs include the use of a single variable for different purposes in disjunct program slices; simulation of a formal parameter using a global variable to which a range of different variables are assigned; and the use of a PRINT-LINE string variable for collecting output from various variables.

The need to avoid pollution is the reason to introduce subtyping for assignments, rather than just type equivalences. In [DM99], we have described a range of experimental data showing the effectiveness of subtyping for dealing with pollution.

Example Figure 5.1 contains a Cobol fragment illustrating various aspects of type inferencing. It starts with a data division containing the declaration of variables. The second part is a procedure division containing statements from which type relations are inferred.

In line 40, variable A00-FILLED is compared to N100, which in line 38 is compared to A00-MAX. This results in an equivalence class between the primitive types of these three variables. Observe that these three variables are also declared with the same picture (in lines 13, 14, and 18).

In line 29, we infer from the assignment that the type of NAME is a subtype of the type of NAME-PART. From line 24, we infer that INITIALS is a subtype of of NAME-PART as well, thus making NAME-PART the common supertype of the other two. Here the three variables are declared with different pictures, namely strings of different lengths. In fact, NAME-PART is a global variable simulating a formal parameter for the R300-COMPOSE-NAM (Cobol does not support the declaration of parameters for procedures). Subtyping takes care that the different sorts of actual parameters used still have different types.
5.3 Concept Analysis

Concept analysis is a mathematical technique that provides a way to identify groupings of items that have common features [GW99]. It starts with a context: a binary table (relation) indicating the features of a given set of items. From that table, the analysis builds up so-called concepts which are maximal sets of items sharing certain features. The relations between all possible concepts in a binary relation can be given using a concise lattice representation: the concept lattice.

Recently, the use of concept analysis has been proposed as a technique for analyzing legacy systems [Sne98]. One of the main applications in this context is deriving (and assessing) the modular structure of legacy software [DK99b, LS97, SR97, ST98]. This is done by deriving a concept lattice from the code based on data usage by procedures or programs. The structure of this lattice reveals a modularization that is (implicitly) available in the code. In [DK99b], we used concept analysis to find groups of record fields that are related in the application domain, and compared it with cluster analysis.

In the remainder of this section we will explain concept analysis in more detail.

---

Table 5.1: The list of items and their features

<table>
<thead>
<tr>
<th>items</th>
<th>features</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>zipcd number-ext number street city prefix initial title name</td>
</tr>
<tr>
<td>1</td>
<td>zipcd number-ext number street city</td>
</tr>
<tr>
<td>2</td>
<td>street</td>
</tr>
<tr>
<td>3</td>
<td>city</td>
</tr>
<tr>
<td>4</td>
<td>prefix initial title name</td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: All concepts identified for Table 5.1.
5.3.1 Basic Notions

We start with a set $\mathcal{M}$ of items, a set $\mathcal{F}$ of features, and a binary relation (table) $T \subseteq \mathcal{M} \times \mathcal{F}$ indicating the features possessed by each item. The triple $(T, \mathcal{M}, \mathcal{F})$ is called the context of the concept analysis. In Table 5.1 the items are the field names, and the features are usage in a given program. We will use this table as example context to explain the analysis.

For a set of items $I \subseteq \mathcal{M}$, we can identify the common features, written $\sigma(I)$, via:

$$\sigma(I) = \{ f \in \mathcal{F} \mid \forall i \in I : (i, f) \in T \}$$

For example, $\sigma(\{ZIPCD, STREET\}) = \{P_4\}$.

Likewise, we define for $F \subseteq \mathcal{F}$ the set of common items, written $\tau(F)$, as:

$$\tau(F) = \{ i \in \mathcal{M} \mid \forall f \in F : (i, f) \in T \}$$

For example, $\tau(\{P_3, P_4\}) = \{STREET\}$.

A concept is a pair $(I, F)$ of items and features such that $F = \sigma(I)$ and $I = \tau(F)$. In other words, a concept is a maximal collection of items sharing common features. In our example,

$$\{(\text{PREFIX}, \text{INITIAL}, \text{TITLE}, \text{NAME}), \{P_1\}\}$$

is the concept of those items having feature $P_1$, i.e., the fields used in program $P_1$. All concepts that can be identified from Table 5.1 are summarized in Table 5.2. The items of a concept are called its extent, and the features its intent.

The concepts of a given table are partially ordered via:

$$(I_1, F_1) \leq (I_2, F_2) \iff (I_1 \subseteq I_2 \iff F_2 \subseteq F_1)$$

As an example, for the concepts shown in Table 5.2, we see that $\bot = c5 \leq c3 \leq c1 \leq c0 = \top$.

This partial order allows us to organize all concepts in a concept lattice, with meet $\wedge$ and join $\vee$ defined as

$$(I_1, F_1) \wedge (I_2, F_2) = (I_1 \cap I_2, \sigma(I_1 \cap I_2))$$

$$(I_1, F_1) \vee (I_2, F_2) = (\tau(F_1 \cap F_2), F_1 \cap F_2)$$

The visualization of the concept lattice shows all concepts, as well as the relationships between them. For our example, the lattice is shown in Figure 5.2.

In such visualizations, the nodes only show the "new" items and features per concept. More formally, a node is labeled with an item $i$ if that node is the smallest
concept with $i$ in its extent, and it is labeled with a feature $f$ if it is the largest concept with $f$ in its intent.

For a thorough study of the foundations of concept analysis we refer the reader to [GW99].

### 5.4 Combine Types and Concepts

In [DK99b] concept analysis was used to find structure in a legacy system. The variables of a Cobol system were considered items, the programs features, and the "variable used in program" property as a relation. Table 5.1 is an example of such a
relation, and Figure 5.2 show the corresponding lattice. This lattice can be seen as a candidate object oriented design of the legacy system. The concepts are individual classes and related concepts can be seen as subclasses or class associations.

The identification of variables in different programs was performed by comparing variable names, and variable declarations. If two variables shared a particular substring they were considered equal. This works well for systems that employ a coding standard which forces similar names for similar variables but fails horribly for systems where variable names are less structured. In this paper this problem is solved by taking the types (as described in Section 5.2) of these variables, and relating them to programs in various ways.

### 5.4.1 Data for Concept Analysis

Before describing the concept experiments performed, first the relations derived from the legacy source will be explained. The four extracted relations are varUsage, typeEquiv, transSubtypeOf and formalParam. varUsage is the relation between a program and the variables that are used in that program. typeEquiv is the relation between a type name (the name of a type equivalence class) and a variable that is of this type. transSubtypeOf is the relation between a type and the transitive closure of all its supertypes, i.e. between two types where the second is in the transitive closure of all the supertypes of the first. formalParam is the relation between a program and the types of its formal parameters. An overview of these relations is given in Table 5.3.

In the remainder of this section the set of all programs, variables, and types in a system will be denoted $P$, $V$, and $T$, respectively.

### 5.4.2 Experiments Performed

**Type Usage**

The first experiment performed is exactly the experiment performed in [DK99b], as described earlier. The type usage per program is taken as the context relation, instead of variable usage. This results in a lattice where the programs that use exactly the same set of types will end up in the same concept, programs that use

<table>
<thead>
<tr>
<th>Relation name</th>
<th>Name of relation element</th>
</tr>
</thead>
<tbody>
<tr>
<td>varUsage</td>
<td>program</td>
</tr>
<tr>
<td>typeEquiv</td>
<td>program</td>
</tr>
<tr>
<td>transSubtypeOf</td>
<td>sub</td>
</tr>
<tr>
<td>formalParam</td>
<td>program</td>
</tr>
</tbody>
</table>

Table 5.3: Derived and inferred relations
5.4 · Combine Types and Concepts

Figure 5.4: Concepts involving relevant programs

less types will end up in a concept below, and programs that use more types will end up in a concept above that concept.

In order to arrive at the type usage concept lattice the varUsage table is taken as a starting point. For each variable, its type is selected from typeEquiv such that the result is a set of relations \( \{(p, t) \in P \times T \mid (p, v) \in \text{varUsage}, (t, v) \in \text{typeEquiv}\} \). Then the types are considered items, and the programs features and the concept analysis is performed. For the example Mortgage system, the resulting concept lattice is shown in Figure 5.3. The list of items and features is not shown for (obvious) lack of space.

Filtering This picture may not be as insightful as we might hope. A way to decrease the complexity of this picture is by filtering out data before performing the
concept analysis. A selection of relevant programs from all programs in a Cobol system can be made as described in [DK98]. Cobol systems typically contain a number of programs that implement low-level utilities such as file I/O, error handling and memory management. These programs can in general be left out of the analysis, particularly when we are only interested in the general structure of the system.

Filtering out insignificant variables is also possible. Typically, certain records in a Cobol system contain all data that has to do with customers (and therefore is probably relevant) while other records may only be used as temporary storage.

Suppose a list of relevant programs is selected and only the data that originated from a certain set of records is deemed interesting. The first step, filtering out the uninteresting programs, is easy. All tuples from varUsage that have an irrelevant program as their program element are simply ignored. Suppose $P_{rel}$ with $P_{rel} \subseteq P$ is the set of relevant programs which is derived in some way. Then all types that are related to the interesting variables need to be determined. Suppose $V_{rel}$ with $V_{rel} \subseteq V$ is the set of all relevant variables. From the relation typeEquiv all types that are related to a relevant variable are selected. If $T_{rel}$ with $T_{rel} \subseteq T$ is the set of all relevant types: $\{ t \in T | (t, v) \in \text{typeEquiv}, v \in V_{rel} \}$ Then the type equivalent variables that are used in the selected relevant programs are selected: $\{ (v, p) \in V_{rel} \times P_{rel} | (v', p) \in \text{varUsage}, (t, v') \in \text{typeEquiv}, t \in T_{rel} \}$

The result of the experiments with filtered data are much more comprehensible than those without filtering, basically because there are less concepts to try to understand. Figure 5.4 shows the concept lattice for the same system as in Figure 5.3, but with irrelevant programs filtered out according to [DK98]. The relevant data are the fields of the two records describing the persistent data in the system.

The lattice in Figure 5.4 contains some unexpected combinations. Concept 7 for instance, contains items that have to do with locations and addresses, but also a birth date. Close inspection reveals that this is not a case of type pollution, but these variables are really used in both program 31(c) (from concept 1) and program 22 (from concept 6). A possible explanation could be that these programs send birthday cards.

It is important to have some way to validate these lattices externally, to perfect the filter set. For our example system, one program implements a utility routine through which a lot of variables are passed, causing one type to contain a remarkable large number of variables. When we filtered out that program, the resulting lattice was much more intuitive.

**Parameter Types**

Experiments have been performed on another concept analysis context; the context that has programs as items and the types of their formal parameter as features. When concept analysis is performed on this data set, all programs that share exactly the same set of parameter-types end up in the same concept. If two programs
share some parameter-types, but not all, the shared parameter types will end up in
the same concept. These will then form an excellent basis for developing an object
oriented view on the system, as the shared types can be seen as the attributes of a
class sharing programs as methods.

In its simplest version the items and features for these concepts are computed
by just taking formalParam and ignoring the subtype relationship.

As was described in Section 5.2, the relation between actual parameter types
and formal parameter types is inferred as a subtype relation. If the subtype rela-
tionship is ignored, then variables can only be identified as having the same type
in different programs, when they are "passed" through a copybook. That is, if a
variable is included in two different programs from the same copybook, it is con-
sidered type equivalent in the two programs. Obviously, this is not the intuition
we have when looking at formal parameters, where we would like to know how
the types used in the calling program propagate to the called program. Therefore,
subtyping is considered as type equivalence when looking at parameter types.

The context for parameter type usage per program while considering super-
types as equivalent is derived as follows:
\{(p, v) \in P \times V \mid (p, t) \in \text{formalParam}, ((t', t) \in \text{transSubtypeOf} \wedge (t', v) \in \text{typeEquiv}) \vee (t, v) \in \text{typeEquiv})\}.

As described in the previous section, data may be filtered on either relevant
programs or relevant data elements. In that case the context is arrived at as fol-
lows: \{(p, v) \in P_{rel} \times V_{rel} \mid (p, t) \in \text{formalParam}, ((t', t) \in \text{transSubtypeOf} \wedge (t', v) \in \text{typeEquiv}) \vee (t, v) \in \text{typeEquiv})\} for some externally determined
value of \(P_{rel}\) and \(V_{rel}\).
An example of a concept lattice showing program as items and the types they use as formal parameters as features (when supertypes are considered type equivalent) filtered for the same set of relevant variables as Figure 5.4 is shown in Figure 5.5.

In this lattice, concept 3 is remarkable, because it contains by far the most programs. This turns out to be caused by the fact that these programs all use "record" as input parameter. Inspection of the source reveals that "record" is a rather large record, and that only some fields of this record are actually used in the programs. It is subject of future work to look at these types of parameters in more detail.

5.5 Refinement of Concepts

When concept analysis is used for analyzing software systems, there will be a point where a user might want to modify an automatically derived concept lattice. For example, consider the applications of concept analysis to remodularization of legacy systems. A maintainer that performs such a task is likely to have knowledge of the system that is being analyzed. Based on that knowledge, he or she might have certain ideas to improve the modularization indicated by the derived lattice by combining or ignoring certain parts of that lattice.

To facilitate the validation of such ideas, we have developed ConceptRefinery, a tool which allows one to manipulate parts of a concept lattice while maintaining its consistency. ConceptRefinery defines a set of generic structure modifying operations on concept lattices, so its use is not only restricted to the application domain of remodularization or reverse engineering. Figure 5.6 shows the application of ConceptRefinery on the data of Table 5.1.

5.5.1 Operations on concept lattices

We allow three kinds of operations on concept lattices. The first is to combine certain items or certain features. When we consider the context of the concept analysis, these operations amount to combining certain rows or columns in the table and recomputing the lattice.

The second operation is to ignore certain items or features. When we consider the analysis context, these operations amount to removing certain rows or columns and recomputing the lattice.

The third operation is combining two concepts. This operation has the following rationale: when we consider concepts as class candidates for an object-oriented (re-)design of a system, the standard concept lattice gives us classes where all methods in a class operate on all data in that class. This is a situation that rarely occurs in a real world OO-design and would result a large number of small classes
that have a lot of dependencies with other classes. The combination of two concepts allows us to escape from this situation.

On the table underlying the lattice the combination of two concepts can be computed by adding all features of the first concept to the items of the second and vice versa.

### 5.5.2 Relation with source

When a concept lattice that was previously derived from a legacy system is manipulated, the relation between that lattice and the code will be weakened:

- Whenever features, items or concepts are combined, the resulting lattice will represent an abstraction of the source system.
- Whenever features or items are ignored, the resulting lattice will represent a part of the source system.

The choice to allow such a weakening of this relation is motivated by the fact that we would rather be able to understand only part of a system than not being able to understand the complete system at all. However, in order for ConceptRefinery to be useful in a real-world maintenance situation, we have to take special care to allow a maintainer to relate the resulting lattice with the one derived directly from the legacy code. This is done by maintaining a concise log of modifications.

### 5.6 Implementation

We have developed a prototype toolset to perform concept analysis experiments. An overview of this toolset is shown in Figure 5.7. The toolset separates source code analysis, computation and presentation. Such a three phase approach makes it easier to adapt to different source languages, to insert specific filters, or to use other ways of presenting the concepts found [DK98, DM98].

In the first phase, a collection of facts is derived from the Cobol sources. For that purpose, we use a parser generated from the Cobol grammar discussed in [DBSV97a]. The parser produces abstract syntax trees that are processed using a Java package which implements the visitor design pattern. The fact extractor is a refinement of this visitor which emits facts at every node of interest (for example, assignments, relational expressions, etc.).

From these facts, we infer types for the variables that are used in the Cobol system. This step uses the Cobol type inferencing tools presented in [DM99]. The derived and inferred facts are stored in a MySQL relational database [YRK99].

In the next phase, a selection of the derived types and facts is made. Such a selection is an SQL queries that results in a table describing items and their features. A number of interesting selections were described in Section 5.4. The
Figure 5.6: Screendump of ConceptRefinery.
results of these selections are stored in a repository. Currently, this is just a file on disk.

In the final phase, the contents of the repository are fed into a concept analysis tool, yielding a concept lattice. We make use of the concept analysis tool that was developed by C. Lindig from the University of Braunschweig. The concept lattice can be visualized using a tool that converts it to input for dot [GKNV93], a system for visualizing graphs. The lattices in Figures 5.2, 5.5, 5.3 and 5.4 were produced this way.

Furthermore, the lattice can be manipulated using ConceptRefinery. This tool allows a user to select items, features or concepts and perform operations on that selection. These operations result in updates of the repository. We distinguish the following manipulations and describe the actions that are carried out on the repository: (1) *combining items or features* is done by merging corresponding columns or rows in the repository; (2) *ignoring items or features* is done by removing corresponding columns or rows in the repository; (3) *combining concepts* is done by adding all features of the first concept to the items of the second and vice versa. The user interface of ConceptRefinery is shown in Figure 5.6. On the left hand side a visualization of the concept lattice is given. The items, features or concepts that need to be modified can be selected in this lattice. The right hand side shows all available operations. ConceptRefinery is implemented in Tcl/Tk [Ous94] and Tcldot: an extension for Tcl/Tk that incorporates the directed graph facilities of dot into Tcl/Tk and provides a set of commands to control those facilities.

---

6The tool "concepts" is available from http://www.cs.tu-bs.de/softech/people/lindig/.
5.7 Related Work

Several methods have been described for modularizing legacy systems. A typical approach is to identify procedures and global variables in the legacy, and to group these together based on attributes such as use of the same global variable, having the same input parameter types, returning the same output type, etc. [CCM96, LW90, OT93, Sch91]. A unifying framework discussing such *subsystem classification techniques* is provided by Lakhotia [Lak97].

Many of these approaches rely on features such as scope rules, return types, and parameter passing, available in languages like Pascal, C, or Fortran. Many data-intensive business programs, however, are written in languages like Cobol that do not have these features. As a consequence, these class extraction approaches have not been applied successfully to Cobol systems [CDDF99]. Other class extraction techniques have been developed specifically with languages like Cobol in mind. They take specific characteristics into account, such as the structure of data definitions, or the close connection with databases [CDDF99, FRS94, NK95]. The interested reader is referred to [DK99b] for more related work on object identification.

*Concept analysis* has been proposed as a technique for analyzing legacy systems. Sneltin [Sne98, Sne00] provides an overview of various applications. Applications in this context include reengineering of software configurations [Sne96], deriving and assessing the modular structure of legacy software [LS97, SR97], object identification [DK99b], and reengineering class hierarchies [ST98].

The extract-query-view approach adopted in our implementation is also used by several other program understanding and architecture extraction tools, such as Ciao [CFKW95], Rigi [WTMS95], PBS [SCHC99], and Dali [KC99].

New in our work is the addition of the combination of concept analysis and type inferencing to the suite of analysis techniques used by such tools. Our own work on type inferencing started with [DM98], where we present the basic theory for Cobol type inferencing, and propose the use of subtyping to deal with pollution. In [DM99], we covered the implementation using Tarski relational algebra, as well as an assessment of the benefits of subtyping for dealing with pollution. Type-based analysis of Cobol, for the purpose of year 2000 analysis, is presented by [EHM+99, RFT99]: both provide a type inference algorithm that splits aggregate structures into smaller units based on assignments between records that cross field boundaries. The interested reader is referred to [DM98, DM99] for more pointers to related work on type inferencing.

5.8 Concluding remarks

In this paper we have shown that the combination of facts derived from legacy source code, together with types inferenced from those facts, forms a solid base
for performing concept analysis to discover structure in legacy systems. This extends and combines our previous work on type inferencing for legacy systems and object identification using concept analysis. We implemented a prototype toolset for performing experiments. From these experiments, we can conclude that the combination of type inference and concept analysis provides more precise results than our previous concept analyses which did not involve types.

The combinations discussed in this paper are the following concept analysis contexts:

1. type usage per program
2. types of parameters per program

The latter analysis appears to be particularly suitable as a starting point for an object oriented redesign of a legacy system.

When performing concept analysis to gain understanding of a legacy system, it proves very helpful if the reengineer is able to manipulate the calculated concepts to match them with his knowledge of the system, or to remove parts he know to be irrelevant. We have implemented ConceptRefinery, a tool that allows a software engineer to consistently perform this kind of modifications while maintaining a relation with both the original calculated concepts, and the legacy source code.

5.8.1 Future work

We would like to extend ConceptRefinery to propose a grouping of concepts to the human engineer to consider when refining the lattice. To this end, we will to experiment with applying cluster analysis algorithms to the concept lattice.

We have discussed two particular concept analysis contexts in this paper. We would like to see whether we could use the results of one of these concept analyses to improve the results of the other. i.e. to take the concept found by looking at the parameter types of programs and somehow use those to mark relevant and irrelevant concepts from the variable usage analysis.

Acknowledgments   The many pleasant discussions we had about this paper with Arie van Deursen are greatly appreciated. We thank Joost Visser for his comments on earlier drafts of this paper.