反编译研究现状及其进展*)

Current Situation and Progress on Decompilation Research

陈凯明　　刘宗田

（合肥工业大学微机所　合肥230052）（上海大学计算机学院　上海200072）

Abstract Research on decompilation is always focused on software reverse engineering. This paper introduces research progress and current situation of decompilation in the newest 10 years, describes its important application field and different research methods, some of them are of practical effect. The decompilation development prospect and difficult problems urgently solved are analyzed in the end.

Keywords Decompilation, Reverse engineering, Research method

一、反编译的定义、作用及其结构

编译器的概念众所周知，但逆编译器的概念却还是很新奇，它允许将所定义的目标代码映射到高级表示。计算字典[1]为编译和逆编译给出了下面的定义：

编译：转换高级语言成目标代码的程序……

逆编译：一种试图…从机器代码转换回到与源程序相似的某种程序的程序。

随着软件技术的不断发展，对现有软件的学习、理解、改造、维护和复用日益变得重要。在不侵犯软件版权或者经授权的情况下，引进软件的消化，吸收和汉化也具有巨大的经济效益。因此，逆编译显得越来越重要，归纳起来，它具有如下一些作用：

①维护和改造拥有使用权的程序。按目前的软件交货方式，软件制造者一般只向用户提供软件的机器代码程序和有关软件使用操作的文档，由于技术保护等原因，通常不提供源代码，这给用户维护和改进程序造成了严重的困难。

②恢复自己开发的程序的源代码的丢失了的部分。这种情况并不罕见，例如意外和事故，保管不善，被离职的雇员携走等，

③探测外来软件，检查外来软件是否剽窃了别的软件中部分，为了国家安全和军事的需要，检查来自他国的软件中是否含有潜在危险的部分。

④获取软件可重用知识与部件，软件重用已经引起软件界的广泛重视，但可重用知识与部件的获取是其中的困难之一，从已有软件中获取可重用知识与部件是解决这一困难的途径之一，对于机器代码程序，如果能将它翻译成高级语言程序，将有利于可重用知识与部件的获取。

⑤在安全关键性的系统中，假设编译器得不到足够信任，因此需要反编译可执行代码，并且由于编译通常不经过优化，用反编译技术进行程序验证也是可能的，此方面的使用见文[2]。

⑥用于代码的质量评价中，例如，发现其中结构怎样，或者满足其它要求的各种各样特征。

⑦为了辅助调试过程，现行的调试器通常都将机器代码反汇编到较高级的汇编语言表示。这有助于获得结构化的高级程序形式的更高级的代码表示。反编译能自动搜索代码中的结构，向软件工程师提供最合适的程序，以便帮助她/他理解代码的操作。

⑧特别是，随着 Internet 上的可免费软件和部件的急剧增加，为用户提供了有利条件。但是，这些软件和构件的安全性是不可回避的严重问题，特别是为了构造安全重要的软件系统，在这种新情况下，反编译系统将会发挥不可替代的重要的作用。

上述都是合法的使用。与其它逆向工程工具一样，反编译也面临是否合法使用的问题。使用的合法与否由使用者的动机决定，不由工具本身决定。国外法律专家对逆向工程工具使用的合法性见文[3]，国内也曾在文[4]中就反编译的合法使用问题阐述过类似观点。

反编译的过程通常如下：

目标语言代码→[反汇编]→[语言和编译器版本识别]→[库函数或类库识别]→[预处理]→中间语言→[数据流分析、控制流分析]→[翻译器]→高级语言

*)本文为教育部博士点基金课题"面向 Web 的组件化 CASE 模型研究"一部分，基金编号为 97035901。"基于知识的逆编译系统 DECLER"获安徽省98年科技进步二等奖。陈凯明　博士生，讲师，研究方向：软件工程，刘宗田　博士生导师，研究员，研究方向：人工智能和软件工程。

二、研究历史及现状

1960年 J. K. Donnelly 和 h. Englander 就使用了反编译术语，并建立了一个将 Univac M-460机器代码翻译成 NELLAC 的实验性反编译器。随后有一些反编译实验系统相继被提出，但由于程序设计语言的飞速发展，这些系统未能付诸实际应用，所应用的技术也已远远落后，由于各种原因，70年代少见有关报道。80年代后，人们针对近代流行的程序设计语言重新展开了反编译技术的研究。1998年美国 Mullane 用模式识别反编译 Microsoft BASIC 的小子集。1993年美国 Digital 公司开发了一个能将 VAX 和 mx 软件转换到 Alpha 机器上的反编译环境。1993年英国 Nuclear Electric plc 开发了一个针对 PROM 代码的反编译系统。1993年英国牛津大学研究能产生 C++子集的反编译器的生成技术。爱尔兰 Limerick 大学自80年代到90年代一直在研究反编译技术，1991年爱尔兰 J. J. O'Gorman 研究了 VAX/VMS 下的 C 语言反编译技术。1995年澳大利亚皇后岛工业大学(Queensland University of Technology)研究开发了 dcc 反编译器，该项研究由澳大利亚研究委员会(ARC)资助。1995年韩国也在开展反编译技术的研究。

在国内，1985年，合肥工业大学曾用手工方法反编译过 UNIX 操作系统。1986年起研究 DUAL 68000机器上的 C 语言反编译系统，获国家自然科学基金资助，开发成功了68000C 反编译系统，获国家机电工业部科技进步二等奖。1988年起开始开展在 IBM PC 系列机上研制 C 语言反编译系统，列为国家七五攻关子课题，开发成功了8086C 反编译系统。1988年开始研究 SUN 工作站上的 C 语言反编译系统，列为863课题。1992年起，利用自筹资金，在8086C 的基础上，开展了商品化的 C 语言反编译系统的研究工作，1995年底完成商品化系统 DECLER V1.0和 V1.1。此外，还有北京信息学院、上海交通大学、中科院计算所和辽宁大学、武汉大学、重庆大学、北京控制工程研究所等，其中上海交通大学在 VAX 机上实现了一个 C 语言反编译系统，北京控制工程研究所在 PC 机上实现一个 C 语言反编译系统，哈尔滨工业大学研究 Turbo C 小模式的反编译系统，该研究属于国防科工委八五攻关课题。

虽然反编译技术研究开始较早，但发表的文献不多，早期的研究并未持续下来，至今很少见真正实用的产品推出。O'Gorman 认为，这可能与以下原因有关：①早期的研究者发现所提出的技术对反编译的实现不十分有效，②认为反编译的应用涉及知识产权问题，一些研究正在秘密进行。

三、研究方法

反编译是高度智能的识别过程，是不完全信息的推理过程。期望源代码的完全复原并且不经修改即能再编译执行，这是不现实的。曾经从事过反编译研究的 P. J. Brown 写道："从机器代码逆向创造源程序，通常是不现实的，因为取出一串机器指令，并使它们与源结构反向相关，是非常复杂的模式识别过程"。N. Wirth 写道："当程序的结构被移去后，例如生成机器代码时由编译器所做的那样，那么识别它的意思实际上是不可能的"。

但是，通过一系列的形式的方法将由源代码生产的机器代码翻译成用户可读的接近于高级语言程序的形式是可能的，并且出现了许多实验性原型系统，有的已付诸实用[5]。在国外，反编译研究的文献近年来逐步增多，研究经费呈现逐年上升的势头。目前，反编译的研究方法主要有四种：

1. 基于文法的方法　这是一种传统且比较实用的方法。它针对特定的高级语言文法和目标语言对，通过语法和类型的枚举，进行反编译研究工作。典型的研究有英国牛津大学的 Bowen，澳大利亚 Queensland 大学的 Cristina Cifuents，爱尔兰 Limerick 大学的 O'Gorman。国内如哈尔滨工业大学赵雷博士研究 Turbo C 小模式的反编译系统，合肥工业大学研究的 Dual 68000以及8086 & MSC 5.0的反编译系统。

2. 基于生成编译器的逻辑程序设计　本方法通常使用一系列的规范描述定义如何将源语言中的结构和变量宣称翻译成机器指令的方式构造一个编译器，这些规范描述能直接变成 Prolog 程序。由于 Prolog 语言的逻辑和回溯特性，改变输入和输出，就可以得到逆编译器。英国剑桥大学的 Bowen 等人正在从事这方面的研究工作。

3. 基于知识的逆编译　逆编译器的生成与编译器和版本有关。不同的编译器及版本所生成的目标代码格式、库函数的形式和特殊结构的目标代码形式都会有些不同，系统提供自定义的模式描述语言，由具有一定逆编译知识的用户根据实际需要增加新的模式，使得系统在使用中能不断地进化。典型工作见合肥工业大学的 DECLER 系统，该系统具有很强的实用性。

4. 逆编译器的编译器　仿照 Yacc，规范描述反编译器的输入和输出，构造能够自动生成逆编译器的编译器，可以大大缩短逆编译器的开发周期，目前，已有这方面研究的零星报道，如针对 Java 语言的逆编译器自动生成研究。

四、应用前景和研究难点

从国内外有关文献中可以看出，近年来逆编译的研究经费逐年增加。从 SOURCER 软件的销售发行情况也可见反编译系统的推广前景。SOURCER 是美国一家公司开发的软件，能将 PC 机器上的可执行程序

反汇编，并附有帮助阅读的注释信息、每年售出数十万套。而反编译系统的功能、作用以及效果远远超过反汇编工具，因此市场前景非常广阔。未来主要应用的领域可能在以下几个方面：

1. 软件的改造和维护。从成本进行分析、有时改造已有软件要比重新研制新软件经济得多。但是由于软件开发商一般不提供高级语言编写的软件源代码或者要价太高、因此使用反编译器将可执行代码翻译成源语言、有助于用户阅读和理解程序，国内已有借助 DECLER[5]成功解决软件汉化所遇到的困难问题。

2. 软件开发环境中的程序调试工具。反编译技术可以应用到软件开发环境中的高级语言的调试过程中。现有的一些 DEBUG 调试器如 Codeview，只能调试带符号表的没有优化的可执行程序。调试反编译生成的中间语言代码比汇编语言方便、比高级语言容易。

3. 安全要求极高的程序验证。在安全性要求极高的软件中，源程序通常不经过编译优化，因此反编译要容易得多。通过将源程序和反编译后的程序相比较、可以进一步确认软件的可靠性。

4. Internet 上应用。学习 Internet 大量免费软件的设计思想和方法。

反编译的文献很难找到，作者认为这与反编译的研究极其困难有关。综合分析近年来国内外有关资料，针对特定机器、特定语言和特定编译版本的反编译技术基本成熟。所研制的反编译在一定条件下能达到实用程度。随着计算机软硬件技术的不断进步、反编译研究面临如下问题：1）复合数据类型恢复，准确率低或根本无法识别；2）通用性和适用性不强；3）反编译器的跨平台、跨机型构造；4）输入代码的规范描述。其中数据结构恢复最困难。

参 考 文 献

1 Illingworth V. Dictionary of Computing, 3rd edn. Oxford University Press, 1990

2 Bowen J P, Stavridou V. Safety-critical systems, formal methods and standards, Software Engineering Journal, 1993. To appear Also issued as a Programming Research Group Technical Report PRG-TR-5-92. 1993

3 John W. Logical Copyright, IEE Review 40 1, 1994. 40～41

4 刘宗田、陈复安. 反编译技术研究现状及面临的问题. 计算机科学、1992，19(6)：55～58

5 刘宗田. DECLER 用户使用手册 合肥工业大学微机所，1995. 3

6 Moore C. Renaissance development. In: Proc. of the Fifth Annual Embedded Systems Conf. Part vol. 2, 1993 257～265

7 Bowen J. From programs to object code and back again using logic programming: compilation and decompilation Journal of Software Maintenance: Research and Practice, 1993, 5(4): 205～234

8 赵磊，王开铸. C 反编译控制流恢复的形式化描述及算法. 计算机学报，1998、21(1)：87～91

9 申例民，等. 基于 CFA 和 DTA 的逆编译方法. 小型微型计算机系统，1998(19)：19～23

（上接第121页）

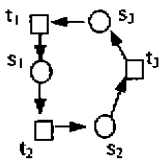

图1 $\sum_1$

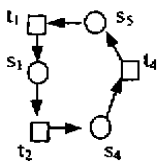

图2 $\sum_2$

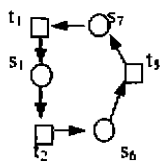

图3 $\sum_3$

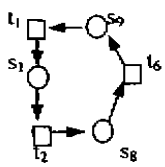

图4 $\sum_4$

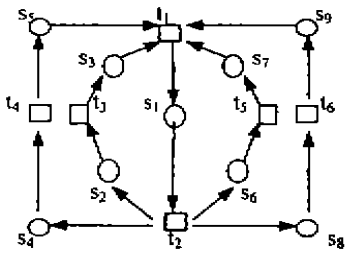

图5 $\sum^M$

参 考 文 献

1 Jiang C J, Wu Z H. Net Operations. Journal of Computer Science and Technology, 1992, 9(4): 333～344

2 蒋昌俊. Petri 网的广义笛积运算. 自动化学报，1993，19(6)，745～748

3 王培良，蒋昌俊. Petri 网的并运算(I). 西北大学学报，1997，27(增)：111～114

4 李孝忠、曹德范、杜玉越，左凤朝. Petri 网的两类广义组合加网. 计算机科学、1999，26(6) 增刊：143～146