

文章编号:1001-9081(2012)04-1060-04

doi:10.3724/SP.J.1087.2012.01060

基于多属性决策的嵌入式操作系统识别技术

张 平*, 蒋烈辉, 刘铁铭, 谢耀滨

(信息工程大学 信息工程学院, 郑州 450002)

(*通信作者电子邮箱 zhangping5312390@163.com)

摘要:针对嵌入式固件逆向解析过程中操作系统类型识别困难的问题,提出了一种基于多属性决策的嵌入式操作系统识别技术。对固件映像中反映出的嵌入式操作系统的多种特征进行综合分析并构建了相关的识别模型,利用向量夹角余弦计算与标准系统之间的相似度,最后阐述了识别的基本思想和具体实现流程。实验结果表明,该方法在某些特征缺失的情况下仍能得到较准确的识别结果。

关键词:嵌入式; 固件; 逆向解析; 操作系统; 多属性决策; 向量夹角余弦; 相似度

中图分类号: TP311 文献标志码:A

Embedded systems recognition based on multi-attribute decision making

ZHANG Ping*, JIANG Lie-hui, LIU Tie-ming, XIE Yao-bin

(Institute of Information Engineering, Information Engineering University, Zhengzhou Henan 450002, China)

Abstract: Concerning the problem that it is difficult to recognize operating system type in embedded firmware reversing analysis, a recognition technology based on Multi-Attribute Decision Making (MADM) was proposed. The paper comprehensively analyzed the multiple features in the firmware, built a recognition model, and calculated the similarity using the vector included angle cosine method. The basic idea of recognition and the concrete realization of the process were described. The experimental results show that this method can get more accurate recognition results in the cases with some features missing.

Key words: embedded; firmware; reverse analysis; operating system; Multi-Attribute Decision Making (MADM); vector included angle cosine; similarity

0 引言

嵌入式操作系统逆向解析是嵌入式固件逆向解析的重要内容和难点之一,通过操作系统的逆向解析,可以实现对固件功能的理解,有利于嵌入式设备的维护和升级^[1]。操作系统的逆向解析必须建立在操作系统类型已知的基础上。然而,在实际工作中,分析人员面向的往往是未知的二进制固件映像,无法确定操作系统类型、版本等相关信息,这样对该操作系统的逆向解析也就无从下手。所以操作系统类型的判别是操作系统逆向解析的前提和基础。

目前针对该项研究较成熟的理论不多,大部分逆向分析人员通常采用关键字匹配的方法,即使用winhex、010editor等二进制分析工具对固件中出现的标识操作系统类型和版本的字符串进行匹配查找,再结合自身经验做出判定。这种方法存在很大局限性,由于特征过于单一,在某些特征信息缺失的情况下无法进行判别,该方法识别效果很不理想,在大多数情况下无法满足用户要求。

针对上述问题,本文提出一种基于多属性决策的嵌入式操作系统识别技术,通过对固件中压缩算法进行识别、对根文件系统进行解析等逆向解析的研究基础上获取操作系统的多种特征。利用基于多属性决策的思想,对固件映像反映出的多种特征进行综合分析,最终得出目标固件映像与标准操作系统的相似程度,并将得出的结果以数据表的形式反馈给用户,作为判定操作系统类型的重要指标。

收稿日期:2011-10-26;修回日期:2011-12-10。

作者简介:张平(1986-),男,辽宁大石桥人,硕士研究生,主要研究方向:嵌入式软件逆向;蒋烈辉(1967-),男,浙江东阳人,教授,博士生导师,主要研究方向:逆向工程;刘铁铭(1977-),男,辽宁锦州人,副教授,主要研究方向:数据库、嵌入式系统;谢耀滨(1981-),男,福建漳州人,讲师,主要研究方向:嵌入式系统。

1 嵌入式操作系统逆向技术

1.1 固件映像剖析

固件是嵌入式设备中操作系统的载体。嵌入式设备不同,固件的结构及特点也不尽相同,通过对Linux、Android等主流操作系统固件进行分析,总结出操作系统固件主要结构如图1所示。



图1 嵌入式操作系统固件主要结构

固件头部主要标识固件编号、厂商等信息;引导程序(Boot Loader)是操作系统内核执行前执行的一小段代码,为操作系统的内核的运行完成必要的初始化工作;内核是操作系统的内核,负责进程调度等核心工作;根文件系统存储系统应用程序以及库文件等系统运行必要的文件。由于受存储空间限制,某些固件的内核和根文件系统以压缩形式存储于固件中,必要时由引导程序解压缩到内存中运行。

1.2 嵌入式操作系统逆向解析模型

嵌入式操作系统逆向解析是针对固件中的操作系统进行

剥离并对其功能模块、系统代码、用户应用程序进行逆向分析的过程。主要步骤有:内核识别与剥离、文件系统识别与剥离、操作系统类型识别、函数分析、代码完整性验证分析等。嵌入式操作系统逆向解析模型如图2所示。

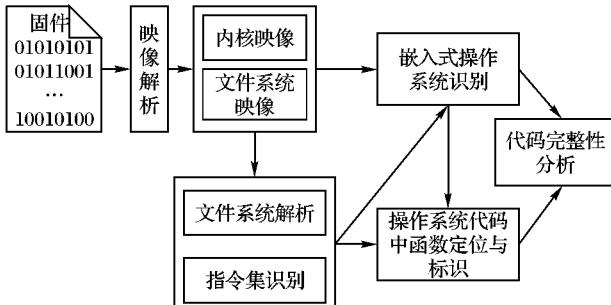


图2 嵌入式操作系统逆向解析模型

2 基于多属性决策的嵌入式操作系统识别

2.1 基本思路

多属性决策问题可以形式化表述为:给定备选方案的集合 $A := (a_1, a_2, \dots, a_n)$, 每个方案具有属性向量 $Attributes := (x_1, x_2, \dots, x_m)$, 决策时根据 A 中每个方案的属性向量对每个方案进行综合评价。在评价时,可以根据属性向量 $Attributes$ 中属性的地位不同分配不同的权值,最终对 A 中所有方案进行综合评价排序,决策者可以根据排序结果做出决策。

基于多属性决策的基本思想,本文所提出的识别方法的基本思路如下:通过对目标固件映像的处理并分析,提取出能反映操作系统类型的多种特征,构造相似度决策矩阵,根据相似度求取算法计算目标固件与特征数据库中备选标准操作系统之间的相似程度,特征数据库中存储不同嵌入式操作系统的特征,它是识别过程的前提和基础。

2.2 相关定义

定义1 操作系统相似度。所谓操作系统相似度,是指通过评估算法计算出来的,反映两个操作系统之间的相似程度的数值。设操作系统 OS_1 和 OS_2 的相似度用 $sim(OS_1, OS_2)$ 表示,其中 $sim(OS_1, OS_2) \in (0, 1)$ 。

定义2 特征属性集。特征属性集是操作系统中能够反映操作系统类型的特征属性的集合。

定义3 操作系统识别决策矩阵:操作系统识别决策矩阵是特征指标集合的汇总,提供了操作系统识别的基本信息,是操作系统识别的根本依据。

2.3 嵌入式操作系统识别特征属性集的选取

通过分析多款嵌入式操作系统源码和系统结构并进行总结与分类,根据操作系统特征属性抽取复杂程度的不同,本文将特征分为简单特征和复杂特征两类。

所谓简单特征是指目标二进制固件映像不经过处理可以直接获得的特征。主要有以下几种。

1)关键字符串。关键字符串主要是指固件映像中存在的某些可以标识操作系统类型的特殊字符串,如固件头标识等,关键字符串可能存在多个。

2)固件映像文件大小。嵌入式设备中,某些嵌入式操作系统固件的映像文件大小是呈一定分布规律的,所以,当固件映像文件的大小符合某种范围时,被认定为符合该特征。

简单特征的提取和匹配实现较为简单,但这种特征存在的几率比较低,所以单纯采用这类特征识别准确度有限,需

要复杂特征才能得到更准确的识别结果。

所谓复杂特征是指需要对二进制固件映像做进一步处理才能获得的某些特征。

1)指令集类型。不同操作系统支持的指令集类型不完全相同,某些特殊的操作系统只在某些特定的指令集下才能运行,即使某些系统支持多种指令集,但支持的范围是有限的。所以固件所使用的指令集类型也可以作为判定操作系统类型的重要标准之一。

2)固件中采用的压缩算法。操作系统支持的压缩算法种类是确定的,所以固件中压缩算法的采取与否和采取压缩算法的种类可以作为识别特征。固件中常采用的压缩算法有 Gzip^[2]、Bzip2^[3] 等。

3)文件系统类型。嵌入式操作系统文件系统类型有多种,如 ext2、ext3、YAFFS2 等。目前最流行的智能手机操作系统 Android 采用的是 YAFFS2 文件系统^[4],而实时操作系统 VxWorks 经常将基于 Flash 的 TFFS 文件系统^[5] 作为根文件系统。

4)文件类型。不同操作系统支持的文件类型是不同的,如 Windows 可执行文件为 PE 文件,而 Linux 下为 ELF 文件。通过对文件系统的还原,可以得到目标操作系统的相关文件,通过判定文件的类型,包括可执行文件的类型及系统库文件的类型,同样可以为操作系统的识别提供必要的证据。

5)操作系统内核类型。操作系统内核类型也是判定操作系统类型的重要特征之一,但不是唯一的特征,多种操作系统基于同一内核的情况是存在的,如嵌入式 Linux 与 Android 系统同样基于 Linux 内核,但 Android 只支持 2.6 以上内核。

2.4 嵌入式操作系统识别相似度决策矩阵

相似度决策矩阵是操作系统识别基本信息的提供者,是相似度计算过程的基础,决策矩阵构造的基本思路为:设特征库中存储的待挑选的操作系统集合为 $X = (x_1, x_2, \dots, x_m)$,其中操作系统 x_i 的特征属性的集合为 $Y_i = (y_{i1}, y_{i2}, \dots, y_{in})$,特征矩阵以 X 为行, Y 为列构造,其中 y_{ij} 表示第 i 个操作系统第 j 个特征属性的值。

决策矩阵中特征属性值的是根据固件映像中提取出的特征与特征库中标准操作系统应有的特征进行匹配的结果,取值范围为 0 到 1。根据匹配的结果的不同相应的属性值设定如下。

1)完全匹配。固件特征符合特征库中特征,如固件的指令集类型是该操作系统支持的指令集类型,特征属性值为 1。

2)部分匹配。该特征专指关键字特征,如果关键字特征有多个,而在固件映像中只出现部分,则将固件映像中出现的关键字特征的数量与库中总数量的比值作为特征属性值。

3)不匹配。固件特征不符合库中的特征,特征属性值为 0。

3 嵌入式操作系统识别流程

3.1 总体流程

识别过程分为 4 个阶段:简单特征抽取、复杂特征提取、相似度计算和综合判定。在决策理论中,简单和复杂特征提取阶段称为决策信息获取阶段,相似度计算与综合评估阶段称为根据决策信息对备选方案排序和择优阶段。识别流程如图 3 所示。

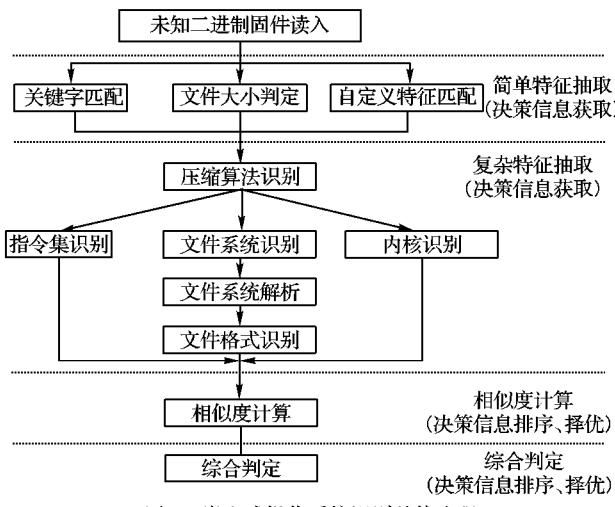


图 3 嵌入式操作系统识别总体流程

3.2 特征抽取

特征抽取过程主要是对简单特征和复杂特征进行抽取的过程,简单特征实现容易,可直接获得,这里不做介绍,目标码指令集识别技术^[6]也有较成熟的技术,能够得到相对准确的结果。这里对压缩算法及根文件系统的识别与解析进行介绍。

1) 固件压缩算法识别。不同压缩算法压缩出的数据表现出的特征不尽相同,但是内部都有一定的存储结构,通过对 gzip 等嵌入式操作系统常采用的压缩算法及压缩文件进行分析,总结出压缩文件通常由三部分组成:压缩文件头,主要标识压缩算法类型(魔数)、压缩文件大小等信息;压缩数据,主要是被压缩数据;压缩文件尾,通常存储压缩数据的校验和等。压缩算法的识别是对一段未知二进制数据中压缩文件进行定位并解压的过程。压缩文件识别与解压缩算法如下。

步骤 1 选择一种需要识别的压缩算法。

步骤 2 匹配该压缩算法的魔数,如果不存在匹配则转步骤 1。

步骤 3 压缩文件头结构分析,如果不符合某压缩算法文件头结构,转步骤 2。

步骤 4 根据压缩文件头定位压缩数据,计算压缩数据大小。

步骤 5 定位压缩文件尾部。

步骤 6 压缩文件转存为相应压缩格式的文件。

步骤 7 调用解压缩程序解压缩。

步骤 8 数据检验,如果解压缩文件不完整或校验错误,转步骤 3;否则结束。

2) 根文件系统识别。文件系统以映像的方式存在于固件中,在固件中都呈现一定的分布规律。固件中文件系统映像的获得主要有两种方法:针对压缩文件系统,经解压缩得到;针对非压缩文件系统,通过定位文件系统的魔数等信息进行定位,再根据超级块的信息进行文件系统尾部的定位,最终提取出剥离出文件系统映像。

3) 根文件系统解析^[7-10]。文件系统解析主要根据文件系统映像物理结构进行解析,文件系统中有两个基本要素:超级块和索引,超级块存储着文件系统的基本信息,索引定义了文件系统内部数据的存储规则。文件系统的解析主要是对超级块进行分析,根据索引关系定位内部文件的存储位置并最终转存的过程。文件系统解析步骤如下。

步骤 1 超级块解析,对文件系统整体信息进行解析。

步骤 2 索引块解析,将所有索引存入待处理队列。

步骤 3 取出队列中待处理索引。

步骤 4 通过索引定位数据块。

步骤 5 对数据块进行转存,如果有多个数据块,则进行数据拼接。

步骤 6 将该索引从队列中删除,队列是否为空,为空则结束;否则转步骤 3。

3.3 相似度计算过程

目前,计算相似度的算法有多种:TOPSIS 法、TOPSIS 夹角度量法、ELECTRE 法、LINMAP 法、OWA 法、AHP 法等^[11]。结合操作系统识别的特点,经过综合考虑,决定采用加权的 TOPSIS 夹角余弦度量法。

算法基本思想是通过相似度决策矩阵的计算得到固件操作系统与特征库中最接近的操作系统。为了反映不同特征的地位不同,引入权重向量为各属性分配相应权重。由于特征属性的值是 0 和 1 之间的值,所以计算过程中决策矩阵不需要归一化处理。相似度计算的具体步骤如下。

步骤 1 得到固件特征属性,结合特征库计算特征属性值,构造出具有 m 款操作系统,每款操作系统有 n 个属性的 m 行 n 列的决策矩阵 A 。

步骤 2 构造权值向量 $w(w_1, w_2, \dots, w_n)$,构成 m 行 n 列加权矩阵 B 。

步骤 3 确定每个属性值的最优值(即特征完全匹配),由于决策矩阵 A 中每列的最优值为 1,所以矩阵 B 中第 i 列的最优值显然为其权值 w_i 。设如果 A 中第 i 行 j 列中元素值为 x_{ij} ,则 B 中第 i 行 j 列元素值为 $x_{ij}w_j$ 。

步骤 4 计算备选的每款操作系统的特征向量 V 与最优向量 V^* 之间的夹角余弦 $sim(V, V^*)$ 。设第 i 款操作系统的 n 维特征向量为 V_i ,其最优的 n 维向量 V_i^* ,则:

$$sim(V_i, V_i^*) = \cos \theta_{(V_i, V_i^*)} = \frac{\sum_{j=1}^n v_{ij} \times v_{ij}^*}{\sqrt{\sum_{j=1}^n v_{ij}^2} \times \sqrt{\sum_{j=1}^n v_{ij}^{*2}}}$$

通过计算特征库中所有备选操作系统进行计算,可以得到相似度 sim 的列表,其中 $sim \in (0, 1)$,该相似度列表即为固件操作系统与各备选操作系统之间的相似度。

3.4 关于权值分配的讨论

相似度计算过程中权值的确定主要根据两点原则:特征的合法程度和特征的重要性。特征的合法程度是指当对固件中的特征进行抽取时对特征的归属无法得到确定的结果而做出的模糊判断,如确定指令集特征时由于用户无法对指令集的归属问题得到确定答案,只能得出与某指令集相似程度的估算结果;特征的重要性是指不同的特征对识别结果的决定程度不同而导致在相似度计算过程中所起作用的大小。

4 实验数据与分析

为了验证本文提出的方法的正确性,选取固件 wrt54g_2_02.7_US_code. bin^[12-13]作为测试数据进行测试。对测试固件进行特征分析,得到的信息如下:

压缩算法为 Gip;大小为 2 885 KB;内核类型为 Linux;指令集为 ARM;文件系统为 Cramfs;文件类型为 Elf;关键字为 Linux, gcc 等。

在得到特征信息表的基础上分配权值向量(1,1,5,2,3,4,4),得到相似度的排名如表 1 所示。

表1 测试固件与不同操作系统的综合相似度

| 排名 | 类型 | 相似度 | 特征属性值 | | | | | | | | |
|----|---------------|---------|----------|--------|----|-------------|------------------|------------------|-------------|--|--|
| | | | 压缩 算法 | 大 小 | 内核 | 指 令 集 | 文 件 系 统 | 文 件 类 型 | 关 键 字 | | |
| 1 | Linux | 0.81342 | 1 | 1 | 5 | 2 | 3 | 4 | 3.2 | | |
| 2 | μ Clinux | 0.62709 | 1 | 0 | 5 | 2 | 0 | 0 | 2.0 | | |
| 3 | Android | 0.53811 | 1 | 1 | 5 | 2 | 0 | 0 | 0.8 | | |
| 4 | VxWorks | 0.12909 | 0 | 1 | 0 | 2 | 0 | 0 | 0.0 | | |
| 5 | μ C/OS-II | 0.09622 | 0 | 0 | 0 | 2 | 0 | 0 | 0.0 | | |

由于篇幅限制,只列出部分备选操作系统。通过相似度排名可知,固件操作系统与备选操作系统中 Linux 系统相似度最高,可以判定该固件所使用的系统为嵌入式 Linux 系统,该判定结果是与实际情况相符合的。实验结果中,Android 与 μ Clinux 与目标固件相似度相对于其他操作系统较高的原因是二者同样基于 Linux 内核,某些特征符合 Linux 特征,但是 Android 操作系统支持的文件系统类型为 YAFFS2,文件类型也不符合;而 μ Clinux 虽然同样基于 Linux 内核,但由于 μ Clinux 系统采用 romfs 作为文件系统,支持的文件格式不是 elf 而是 flat 格式^[14]。因此,二者相似度相对于 Linux 较低。笔者对其他类型操作系统固件进行测试,得到类似结果,验证了本文提出方法的正确性。

5 结语

本文所提出的基于多属性决策的嵌入式操作系统识别技术将多属性决策的思想应用于固件代码逆向解析中,解决了嵌入式操作系统识别的问题。对嵌入式操作系统在固件映像中体现出的多种特征进行综合分析评估,得到固件中的操作系统与各备选操作系统之间的接近程度。实验数据证明,该方法能够较准确地识别出嵌入式固件操作系统的类型,取得

良好效果。下一步的研究方向为:寻找更多的反映操作系统类型的特征属性,进一步提高识别准确度。

参考文献:

- [1] EILAM E. Reversing: Secrets of reverse engineering[M]. Trademarks: Wiley Publishing, Inc, 2005.
- [2] The gzip home page[EB/OL]. [2011-07-27]. <http://www.gzip.org>.
- [3] Bzip2[EB/OL]. [2011-06-10]. <http://www.bzip.org/1.0.5/bzip2-manual-1.0.5.html>.
- [4] 刘敏. 移动终端的 Android 移植与应用程序设计[D]. 西安: 西安电子科技大学, 2011.
- [5] 梅佳希. 嵌入式 VxWorks 下 Flash 文件系统的研究与实现[D]. 武汉: 华中科技大学, 2008.
- [6] 蒋烈辉. 固件代码逆向分析研究与系统设计[D]. 郑州: 信息工程大学, 2007.
- [7] 彭晓曦. 嵌入式 Linux 下文件系统的研究与实现[D]. 成都: 电子科技大学, 2007.
- [8] 顾喜梅. 文件系统及磁盘管理实现机制深入研究[D]. 南京: 南京航空航天大学, 2002.
- [9] BREEUWSMA M, de JONCH M, KLAVER C, et al. Forensic data recovery from flash memory[J]. Small Scale Digital Device Forensics Journal, 2007, 1(1): 1-16.
- [10] BREEUWSMA I M F. Forensic imaging of embedded systems using JTAC[J]. Digital Investigation, 2006, 3(1): 32-42.
- [11] 徐泽水. 几类多属性决策方法研究[D]. 南京: 东南大学, 2002.
- [12] Linksys WRT54G series[EB/OL]. [2011-10-29]. http://en.wikipedia.org/wiki/Linksys_WRT54G_series.
- [13] WRTouters. [EB/OL]. [2011-10-01]. <http://wrt54g.net/>.
- [14] μ Clinux 小型化的做法-可执行文件格式[EB/OL]. [2008-08-02]. <http://linux.chinaunix.net/techdoc/system/2008/08/02/1022247.shtml>.

(上接第 1059 页)

$$CW = \begin{bmatrix} 0 & -0.09\% & -3.86\% & -5.80\% & 0 \\ 0 & -9.28\% & -1.17\% & -0.91\% & 0 \\ -1.66\% & 0 & 0 & 0 & 0 \\ 1.70\% & 0 & 0 & 0 & -1.38\% \\ 0 & 0 & 0 & 0 & -2.87\% \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & -0.03\% \\ 0 & -2.35\% & -1.20\% & -1.30\% & -10.30\% \\ 0 & -8.08\% & -10.13\% & 16.70\% & 0 \\ 9.02\% & 0 & 0 & 0 & 0 \\ -0.19\% & 0 & 0 & 0 & 0.01\% \end{bmatrix}$$

3 结语

本文针对 BP 神经网络的缺陷进行了改进。增加了动量项;采取了自适应调节学习率;引入了陡度因子跳出局部最小值;采用遗传算法优化初始权重和阈值;采取可变隐层节点的办法,进行纵向比较总误差的均方差,选取最优隐层节点数;采用不同训练函数,横向比较网络训练结果精度和速度,选取最优训练函数。算例表明了算法的计算精度和科学性。

本文将改进的 BP 网络用于指挥防护工程围岩自稳能力评估中。结果表明,当样本数据准确获得后,评估网络的效率就会很高,能够节约大量人力资源。该模型在指挥防护工程围岩自稳能力评估中有一定的推广价值。对于模型参数的改进、精度的提高,是下一步研究的重点。

参考文献:

- [1] HECHT-NIELSEN R. Theory of the back propagation neural network[EB/OL]. [2011-05-10]. <http://s112088960.onlinehome.us/annProjects/Research%20Paper%20Library/backPropTheory.pdf>.
- [2] 葛哲学, 孙志强. 神经网络理论与 Matlab R2007 实现[M]. 北京: 电子工业出版社, 2007: 46-47.
- [3] 柳益君, 吴访升, 蒋红芬, 等. 基于 GA-BP 神经网络的环境质量评估方法[J]. 计算机仿真, 2010, 27(7): 121-124.
- [4] 陈刚, 何政伟, 杨斌, 等. 遗传 BP 神经网络在泥石流危险性评价中的应用[J]. 黄金, 2010, 46(3): 228-231.
- [5] SHAN LI, ZHANG JIONG, SUN ZENG-XIAN. Prediction of cyclosporine a blood concentration by genetic algorithm and BP neural network [C]// Bioinformatics and Biomedical Engineering. [S. l.]: IEEE 2008: 729-732.
- [6] GUPTA C D. Back propagation neural network method of solution of normal fat dipole and truncated conical grounded monopole and optimization by genetic algorithm [C]// Antenna Theory and Techniques. Sevastopol, Ukraine: IEEE, 2007: 208-210.
- [7] 段林娣, 宋成辉. 应用 BP 神经网络进行隧道围岩快速分级[J]. 中国安全科学学报, 2010, 6(2): 58-67.
- [8] 付玉华. 露天转地下开采岩体稳定性及岩层移动规律研究[D]. 长沙: 中南大学, 2010.
- [9] 岳万英, 周培根, 霍恩俊, 等. 军事工程百科辞典[M]. 北京: 解放军出版社, 2003: 248-249.