文章编号:1007-5321(2017)05-0129-06 DOI:10.13190/j.jbupt.2016-252

基于 E-OEM 的 SWF 半结构化模型建立

赵海英[1], 谭 欣[1], 王亮亮[2]
(1. 北京邮电大学 数字媒体与艺术设计学院, 北京 100876;
2. 新疆师范高等专科学校 新疆教育学院, 乌鲁木齐 830043)

摘要: 为了揭示 SWF 文件格式的隐含属性,基于对象交换模型(OEM)的构建方式,提出了一种增强半结构化模型 E-OEM,可对 SWF 文件格式进行描述和存储. 采用 OEM 进行 SWF 文件格式的描述;对 OEM 描述模型进行改进,采用头尾分类、文件体聚类的思想将相同类别标签作为一类;通过引入 Huffman 编码,实现 E-OEM 具有描述和存储的功能. 随机选择百例不同源文件进行 E-OEM 建模,仿真实验结果表明,所提模型不仅可以将隐含属性显性表示,同时提高了具有高标签重复率的文件存储效率,证实了模型的有效性.

关 键 词: SWF 文件格式; 增强对象交换模型; Huffman 编码; 半结构化模型
中图分类号: TN911.22 **文献标志码**: A

A Semi-Structure Model of SWF Based on E-OEM

ZHAO Hai-ying[1], TAN Xin[1], WANG Liang-liang[2]
(1. School of Digital Media & Design Arts, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Xinjiang Education Institude, Xinjiang Teacher's College, Urumchi 830043, China)

Abstract: It's difficult to reveal hidden relationships of the shockwave flash (SWF). The article proposed an enhancement object exchange model (E-OEM) based on object exchange model (OEM). This model can describe and store SWF file format. First, SWF file format is described though OEM. Then, OEM descriptive model is improved with classifying the head and the end, clustering the body. E-OEM could realize the function of description and storage through Huffman coding. Also, 100 different source files are used to build semi-structure model. It is shown that E-OEM has a good impact on the relationship dominance, and can implement effective retrieval of high tag repetitive rate, and the storage rate is improved. It is shown that the model is valid.

Key words: shockwave flash file format; enhancement object exchange model; Huffman coding; semi-structured model

SWF(shockwave flash)是动画设计软件 Flash 的专属格式,支持点阵图形与矢量,具有缩放不失真、文件体积小等特点,广泛应用于网页制作、动画设计等领域中. 非结构化的 SWF 文件,结构简单,无法找到各个属性的对应关系,需要依据 SWF 文件说明先进行数据分析,大大降低了处理效率.

半结构化数据代指那些结构隐含或不严谨的数据类型,全世界超过 80% 的数据类型都是以半结构化形式存储的[1]. 它可以很好地解决结构化数据的严格约束与非结构化结构混乱的问题.

笔者基于对象交换模型(OEM, object exchange model)的优点与不足,提出了增强 OEM(E-OEM,

收稿日期: 2016-10-08
基金项目: 国家科技支撑计划项目(2014BAH13F02)
作者简介: 赵海英(1972—), 女, 副教授, E-mail: zhy.yn@163.com.

enhancement OEM)并进行相关研究.

1 相关工作

Zhu[2]对 Flash 动画进行文件结构与上下文结构挖掘,构建了 3 种分类模型实现 Flash 动画的语义特征信息分析. Fouche 等[3]采用数据插入技术,将隐含数据替换为不规则的二进制数,保证了数据的隐秘性. Zhang 等[4]将传统的地理信息系统数据通过 Flash 软件的组建分离,转化为 SWF 文件,具有更高的传输速度. 王岳平[5]采用 SWFmill 对电影文件的内部存储结构进行分析,用于电影的场景中.

迄今为止,常用的半结构化描述方式有 OEM 描述模型与 XML 描述模型. OEM 是斯坦福大学 Serge Abiteboul 教授所提出的一种自描述的数据模型,可以很好地对半结构化数据进行描述与设计[6]. 曹文仙等[7]将整个数据库构建为一个 OEM,采取半结构化数据查询重写桶算法的思想,降低了算法代价. 黄洪等[8]在分析出传统图形用户开发存在的问题后,提出了一种基于 XML 的图形用户界面描述方法,实现了图形用户界面定义与具体程序设计语言和开发平台的无关性. 张富等[9]提出了一个模糊 XML 模型,实现了模糊 XML 模型到知识库间的转化. 孙宏伟[10]提出了 XML 与关系数据库的三层双向集成技术,实现了 XML 与关系数据库之间的数据转换. XML 相比于 OEM 而言,更为灵活,但容易造成大量冗余,在 SWF 格式分析中,采取 OEM 的改进方式更适合 SWF 短小精悍的特点.

在动画及视频的存储方面,张敏[11]利用数字图像处理技术与视频分割技术,为用户提供更精确、生动、全面的教育 Flash 资源,推动教育数字化的发展;Deng 等[12]在 SWF 文件内部读取的字节数组部分加入水印标签来保护版权,防止反编译等有效攻击;陈爱东[13]结合 XML 文件格式特点,实现了 SWF 向 XML 文件的转换,并构建了 Flash 动画检索系统;廖建平[14]根据 SWF 解析规则设计解析框架,实现了文件头与文件体的解析,并采用 XML 描述方式进行形式化描述;倪应华等[15]详细地分析了 SWF 的文件头与标签,并设计解析程序实现了 SWF 动画环境信息、动画元素信息以及控制信息的获取.

2 可描述可存储的半结构化模型

2.1 基于 OEM 的描述模型

SWF 文件格式分为 3 个部分:文件头、文件体,文件尾. 文件头用于确定文件基本信息;文件体描述了文件的细节内容;文件尾仅用作简单的停止. 在 SWF 半结构化模型建立中,引入 OEM 进行描述. OEM 以对象作为实体,每个对象可由一个四元组(oid, label, type, value)进行表示[6]. 笔者选择较为常用的属性进行半结构化描述,构建基于 OEM 的半结构化 SWF 描述模型. 图 1 是源文件 abc.swf 的 OEM 模型. 在建立模型后,需要引入一些 OEM 的相关概念对 SWF 文件格式进行相关描述.

定义 1 一个简单的路径表达式是以点为间隔的标签描述(label)序列,记作 $se = l_1, l_2, l_3, \cdots, l_n$.

定义 2 频繁简单路径表达式 fpe:pe 出现在 OEM 中的次数称为 pe 的支持度计数 sup(pe),来去掉出现频度较低的简单路径表达式. 当 sup(pe) ≥ min_sup(用户自己设立的最低支持度)时,pe 称为频繁 pe,记作 fpe.

由定义可见,fpe 与 dp 相比可以更好地反映 SWF 的一般结构,采用模式抽取的方式[16-18],对图 1 的 OEM 进行深度优先遍历,找到所有频繁的最大简单路径表达式,用 mpe 表示.

在构建 OEM 描述模型时,对遍历得到的最大简单路径表达式 mpe 添加层次标记,从而生成最大简单路径表达式集合,记作 M. M 中包含了 OEM 图中所有存在的路径,其他任何一个路径集均为它的子集. 在图 1 中,经过 OEM 的优先遍历后,可以生成层次结构的最大简单路径表达式集 $M = \{\text{head}_{1-1}.\text{sign}_{2-1}.\ \text{compress}_{3-1}.\ \text{first}_{4-1},\ \text{head}_{1-1}.\ \text{sign}_{2-1}.\ \text{compress}_{3-1}.\ \text{sec ond}_{4-2}\cdots\}$. OEM 构建方式可以构建出完备的 M 集合,但需要较多的存储空间,同时检索效率较低,因此提出一种基于 OEM 改进的增强半结构化模型——E-OEM,该模型可以实现描述与存储的功能.

2.2 E-OEM 半结构化模型建立

对上述建模所遇到的问题,采取改进措施.

首先,对整个 SWF 文件进行遍历,标记出各个标签的位置信息作为聚类的类内标识(location)并获得每个标签的 ID 号. 在 SWF 文件中,文件头、文件尾属性固定,采用决策树方式对其进行分类. 文件体部分长度不固定,采用 K-means 算法进行聚类. 聚类个数参照 SWF7.0 说明的标签数量, ID 号相同的标签为一类. 将同类标签压入堆栈进行存储,按照 location 进行编号,避免查询时的混乱. 对相似标签内容,根据需要进行压缩、整合,该方法极大程度

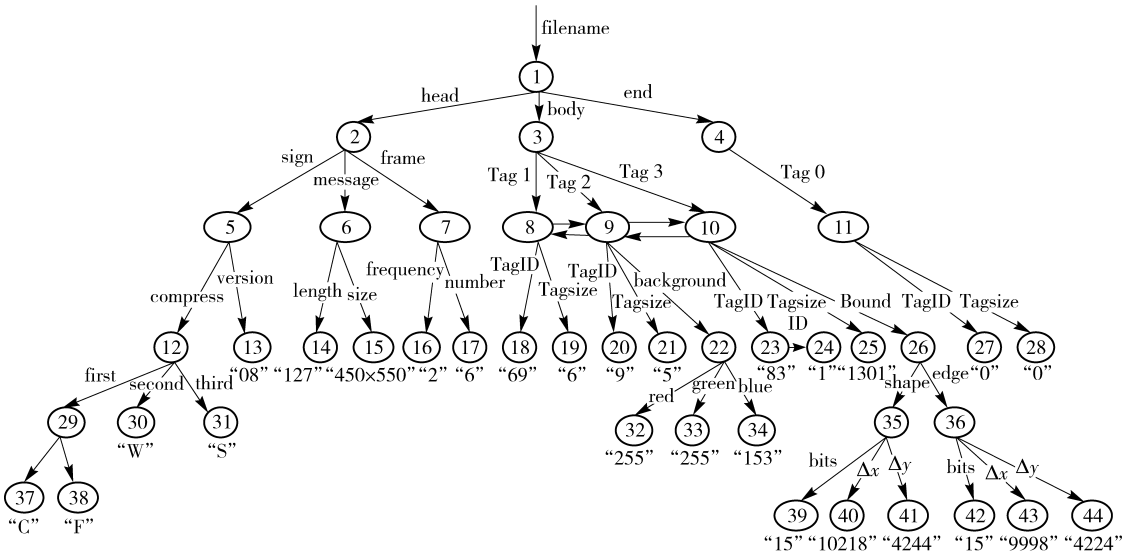


图 1　基于 OEM 的半结构化 SWF 描述模型

上解决了存储过程中冗余量较大的问题，但导致最大简单路径表达式扩大，检索效率降低.

采用 Huffman 编码方式进行存储，对存入堆栈标签的数量进行统计，出现频度最高的标签具有最短的编码位数，逐级递增. 对于一个 SWF 文件而言，一般 4 ~ 5 位编码即可满足视频播放的要求. 采用此种编码方式有效地压缩了存储空间，具有实际意义. 为方便描述，对 E-OEM 进行如下定义.

定义 3　一个 SWF 文件可以构建为一棵以 SWF 文件标签 ID 为节点的 Huffman 树 T，用三元组表示 $T=(\mathrm{ID},R,W)$. ID 代表 SWF 文件标签的 ID 号；R(relation) 是一个属性集，在其中包含了相同 ID 下所有的不同属性、位置以及属性值；W(weigh) 代表 Huffman 的权重.

定义 4　R 属性值较多，根据 ID 的不同，将其构建成单独的一系列子树 $P=\{P_1,P_2,P_3,\cdots,P_n\}$，可以看出 $P\subseteq T$.

定义 5　$\forall P_i\in P,\{L,A,C\}\subseteq P_i$，$L$(location)代表在 SWF 文件中，该标签所在的位置信息；A(attribute)代表该标签所含的属性信息；C(correlation)代表一个关联关系，包含 3 种属性 (U,V)，$(U,V)\in R$. 当 $U,V\in\mathbf{N}$ 时，下列结果可以成立：①U 为父对象，V 为子对象；②U 为父属性，V 为子属性；③U 为原子属性，V 为属性值.

定义 6　(相似度判定) P_i 是一个 R 的子树，对其用三元组进行表示. 不同标签间不具备可比性，因此针对同一个 R 内的所有子树进行分析，$R_A=\{A_1,A_2,\cdots,A_n\}$，子权重按照均匀分布设置为 $W_i=\dfrac{1}{n}$，$\sum_{i=1}^{n}W_i=1$. 同时，构建出一个线性无关的集合 $R'_A=\{A'_1,A'_2,\cdots,A'_k\}$，$\forall A'_p\in R'_A$，$p\leqslant k$，不存在 $A'_n=k_1A'_1+\cdots+k_pA'_p+\cdots+k_nA'_n$，其中 $p\leqslant n$. 在此规定，相似度(similarity)可表示为

$$S=\frac{\sum_{i=1}^{k-1}W'_i}{\sum_{p=1}^{n-1}W_i}=\frac{k-1}{n-1},\ S\in[0,1] \tag{1}$$

当 S 趋向于 0 时，相似度较高；S 趋向于 1 时，相似度较低. 当 $(n-1)(k-1)=0$ 时，代表该类中仅存在一个标签，将 S 置为 ONLY. Ni 等[19]对 SWF 文件进行分析，得出同一 ID 的相似度 S 较高，做聚类操作是有意义的.

3　仿真结果

根据 E-OEM 描述存储的双特性，将 E-OEM 结构模型应用于 SWF 解析中，做出了 SWF 的解析器，实现 SWF 文件属性显性化. 该解析器的技术路线包括以下几步：①借助 E-OEM 对 SWF 的各项数据进行信息重构，分类列出相关信息；②根据找寻最大简单路径表达式的方法，确定对象、属性、属性值之间的相互关系，并确认其中的隐含属性；③采用 Huffman 树状存储方式对相同 ID 进行存储，节约存储空间，并提高检索效率；④构建可视化的 SWF 解析器来对模型进行验证.

对 SWF 文件采用 E-OEM 进行描述，构建可视

化的 SWF 解析器可以明确其内部的结构关系，如图 2 所示.

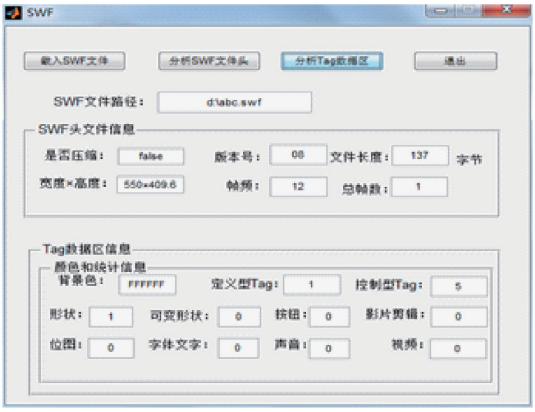


图 2 SWF 半结构化文件解析

由图 2 可以看出，半结构化解析的 SWF 文件具备显示隐含属性的能力. 此框架可以通过标签分类完成 SWF 文件的分析，在文件头部分进行具体阐释，文件体的标签数量众多，采用数值统计的方法，将其代表的主属性表示出来，对整体影响较小的属性进行了适当忽略. 可以看出，半结构化模型相比于非结构化模型具有更清晰的层次性、逻辑性与描述性.

同时，根据 SWF 的编码规则，对图 1 的源文件进行半结构化描述，构建 E-OEM 描述存储结构示意图，如图 3 所示.

如图 3 所示，通过对图 1 源文件标签的遍历、统计构建 SWF 文件标签分布图（图 3(a)），利用式(1)计算源文件各标签的相似度 S（图 3(b)），进而构建图 1 源文件的 E-OEM（图 3(c)）. 相比图 1 的 OEM，图 3(c)所示的 E-OEM 模型更加短小精炼.

相比图 3 源文件，图 4 是一个较为复杂的 SWF 动画源文件 motion. swf 的模型构建过程. 采用相同方式进行模型构建.

标签出现的频率称为标签重复率. Huffman 编码将标签重复率高的标签赋予短的编码长度，减少存储空间的浪费. 实验中，使用最高的标签重复率（最大标签重复率）判定压缩率与标签频度的关系.

在检索方面，通过增加最大简单路径表达式权重，尽可能缩小检索范围，在高频 ID 中表现明显. 引入相似度 S，将相似程度很高的标签归类，有效减少冗余量. 采用子树的方式对复杂属性进行二度分类，具有更好的适应性. 与 OEM 描述模型相比，E-OEM拥有更快的检索效率，更优的存储空间.

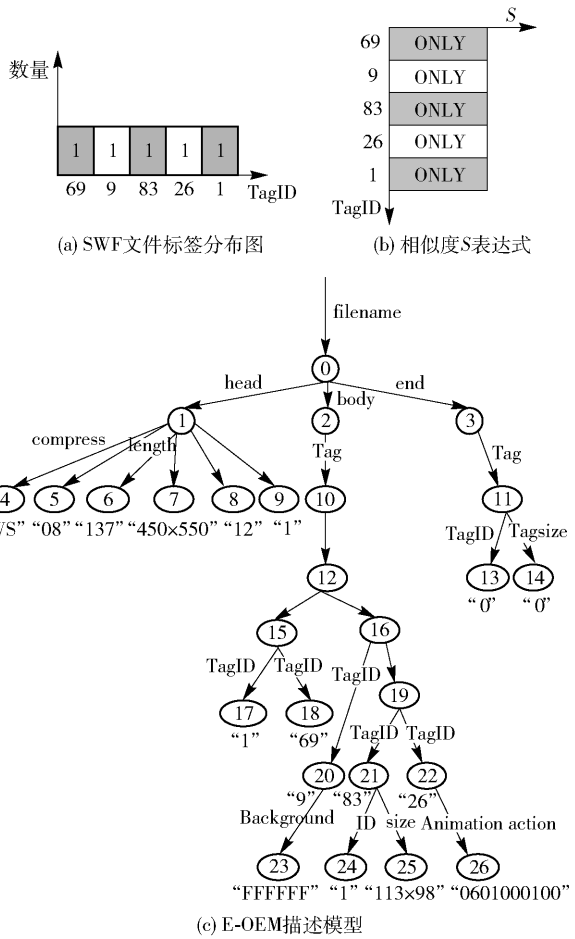


图 3 构建图 1 源文件的半结构化描述存储模型

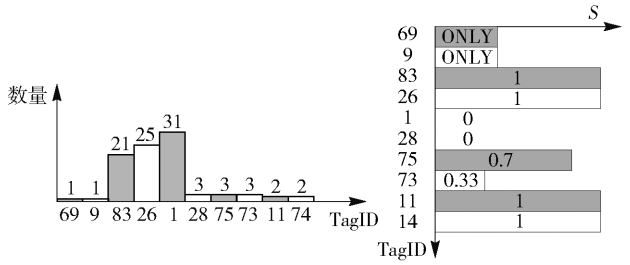
在实际的数据解析中，文件涵盖的信息会比实验中列举的更多. 在标签类别多的情况下，更能体现出 E-OEM 的优越性. 图 5 选取了 100 例 SWF 文件进行测试，采用 SWF 协议的压缩解压方法与模型方法的理论值进行对比.

由图 5 可以直观看出，同在 ID 层进行压缩率对比，E-OEM 压缩率在最大标签重复率较高时，具有更好的压缩性能，如表 1 所示.

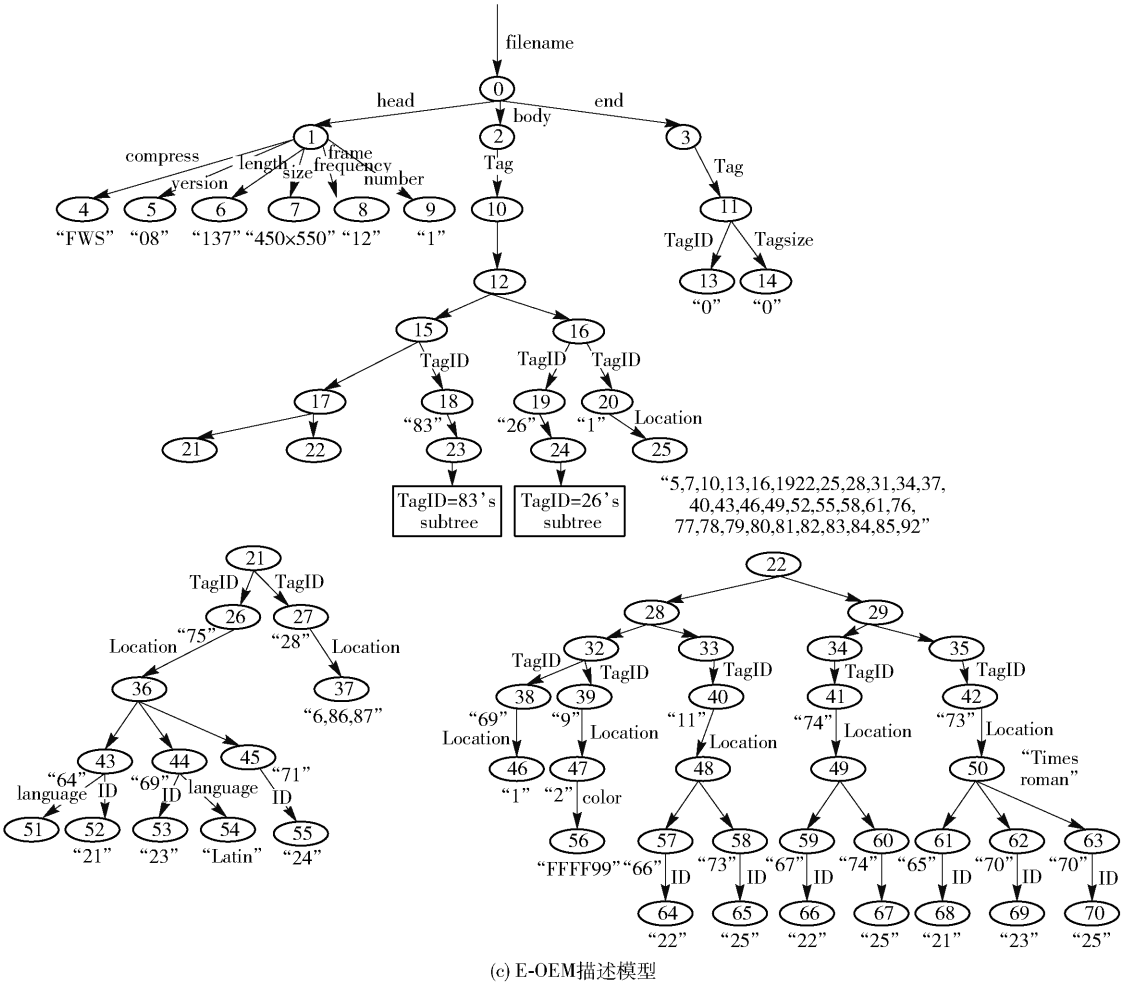
表 1 部分 SWF 压缩率数据对比 %

文件	SWF 协议压缩率	E-OEM 压缩率	最大标签重复率
1	84.63	74.85	38.18
2	75.89	59.95	41.40
3	82.22	67.69	39.41
4	94.43	71.48	30.46
5	85.46	41.74	65.08

因此，在模型存储方面，文件的最大标签重复率将会影响模型的存储效率. 在检索方面，经过全局遍历后，对于特定的检索要求，E-OEM 可以达到处



(a) SWF文件标签分布图　　(b) 相似度S表达式



(c) E-OEM描述模型

图 4　较复杂源文件的半结构化描述存储

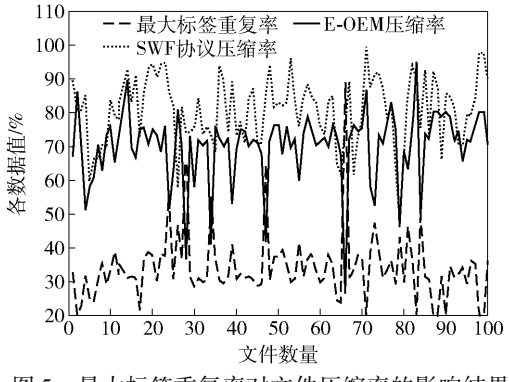


图 5　最大标签重复率对文件压缩率的影响结果

理的快速性. 总体而言, E-OEM 在存储方面可以基本实现模型功能,具有普适性和广泛性.

4　结束语

笔者提出一种可描述可存储的半结构化模型 E-OEM,该模型可用于提高存储效率;同时,引入 Huffman 编码与相似度 S,提高检索效率. 在 SWF 文件解析中,实现了标签隐含属性显性化,一定程度上提高了存储率与检索率,对非结构化数据具有推广意义,仿真结果证明了模型的有效性.

致谢 感谢移动媒体与文化计算北京市重点实验室的支持

参考文献:

[1] 孙涛. 面向半结构化数据的数据模型和数据挖掘方法研究[D]. 吉林: 吉林大学, 2010.

[2] Zhu Xiaowei. Research on automatic classification technology of flash animations based on content analysis[J]. Journal of Multimedia, 2013, 8(6): 693-698.

[3] Fouche M A, Olivier M. Steganographic techniques for hiding data in swf files[C]// IFIP International Conference on Digital Forensics. Berlin Heidelberg: Springer, 2011: 245-255.

[4] Zhang Jinqun, Zhu Yunqiang, Wang Juanle, et al. Flash based WebGIS system and its application in monitoring and evaluating China's regional development[J]. International Journal of Digital Content Technology and Its Applications, 2011, 5: 285-295.

[5] 王岳平. Flash 电影的视觉场景和图形图像特征研究[D]. 济南: 山东师范大学, 2013.

[6] 杨学伟. 基于 OEM 模型的半结构化数据模式抽取算法研究[D]. 北京: 中国石油大学, 2011.

[7] 曹文仙, 赵雪岩, 李建成, 等. 半结构化数据 OEM 图应用[J]. 西安工程科技学院学报, 2007, 01: 92-95.
Cao Wenxian, Zhao Xueyan, Li Jiancheng, et al. Application of the semi structured data OEM diagram[J]. Journal of Xi'an University of Engineering Science and Technology, 2007, 01: 92-95.

[8] 黄洪, 林辉, 王奔. 一种图形用户界面的 XML 描述方法与工具开发[J]. 计算机应用与软件, 2011, 10: 198-202.
Huang Hong, Lin Hui, Wang Ben. A GUI XML description method and tool development[J]. Computer Applications and Software, 2011, 10: 198-202.

[9] 张富, 严丽, 马宗民, 等. 基于模糊描述逻辑的模糊 XML 模型的表示与推理[J]. 计算机学报, 2011, 08: 1437-1451.
Zhang Fu, Yan Li, Ma Zongmin, et al. Representation and reasonng of fuzzy XML model with fuzzy description logic[J]. Chinese Journal of Computers, 2011(8): 1437-1451.

[10] 孙宏伟. XML 与 RDB 的多层次双向数据集成技术研究[D]. 西安: 西北工业大学, 2003.

[11] 张敏. Flash 组成元素的视觉特征研究[D]. 济南: 山东师范大学, 2011.

[12] Deng Hua, Zhang Jifu, Chai Xiaoli. The design and implementation of flash animation watermarking[C]// 2014 IEEE Workshop on Electronics, Computer and Applications. [S.l.]: IEEE, 2014: 489-491.

[13] 陈爱东. Flash 动画的内容提取与描述模型研究[D]. 济南: 山东师范大学, 2010.

[14] 廖建平. SWF 动画内容解析与 XML 表达自动阅卷研究[D]. 上海: 华东师范大学, 2009.

[15] 倪应华, 金炳尧. SWF 矢量动画解析框架设计[J]. 计算机系统应用, 2010, 19(3): 202-205.
Ni Yinghua, Jin Bingyao. Design of an analytical framework with SWF vector graphics[J]. Computer Systems & Applications, 2010, 19(3): 202-205.

[16] 鲁明羽, 陆玉昌. 基于 OEM 模型的半结构化数据的模式抽取[J]. 清华大学学报(自然科学版), 2004(9): 1264-1267.
Lu Mingyu, Lu Yuchang. OEM-based schema extraction of semi-structured data[J]. Journal of Tsinghua University (Science and Technology), 2004(9): 1264-1267.

[17] 杨学伟. 基于 OEM 模型的半结构化数据模式抽取算法研究[D]. 北京: 中国石油大学, 2011.

[18] 李颖, 张晓贤, 孙佳慧. 基于频繁模式半结构化数据的模式抽取[J]. 吉林大学学报(信息科学版), 2012(5): 540-543.
Li Ying, Zhang Xiaoxian, Sun Jiahui. Semi-structured data model extraction based on frequent patterns[J]. Journal of Jilin University (Information Science Edition), 2012(5): 540-543.

[19] Ni Yinghua, Yuan Liyong, Ma Yongjin, et al. Research on content retrieval of flash animation based on XML[C]// 2010 3rd IEEE International Conference on Ubimedia Computing (U-Media). [S.l.]: IEEE, 2010: 64-67.