Inline Text Wrapping Picture

北京邮电大学

硕士研究生学位论文开题报告

学 号: 2013110543

姓 名: 王忠珂

学 院: 计算机学院

专业(领域): 信息安全

研究方向: 网络安全技术与应用

导师姓名: 胡正名

攻 读 学 位: 工学硕士

2014年12月4日

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 基于密度峰值算法的可移植可执行文件病毒聚类分析 | | |
| 选题来源 | 企、事业单位委托项目 | 论文类型 | 应用研究 |
| 开题日期 | 2014-12-07 | 开题地点 | 北京邮电大学 |
| **一、立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）（不少于800字）**  1．研究目的与意义  随着网络和计算机应用的普及，软件程序已深入人们的生活，利用软件系统漏洞的恶意软件潜入网络和计算机系统，盗取或破坏他人信息的事件层出不穷而且每年都在急速增加。  计算机病毒在《中华人民共和国计算机信息系统安全保护条例》中被明确定义，病毒指“编制者在计算机程序中插入的破坏计算机功能或者破坏数据，影响计算机使用并且能够自我复制的一组计算机指令或者恶意程序代码”。因为恶意攻击能够造成巨大的经济损失，所以我们认为成功检测计算机病毒成为计算机安全行业的重中之重。  当今最主流和成熟的反病毒技术还是传统的特征码扫描技术。一个新的病毒从危害用户到反病毒软件捕获其特征码并将之查杀有一个时间差，即使反应足够迅速，但是加上病毒变种的简单迅速和用户条件限制，通过更新特征码库来应对每个未知的病毒就显得力不从心，且会使病毒特征码库无比的庞大，降低扫描效率。  另一方面，对于反病毒技术人员来说，如何将精力集中在大量已有病毒的变种中的新型未知病毒上，对于提高病毒响应效率也非常重要，因此，如何实现对病毒的分类检测，从而将真正的未知病毒和已有病毒分别归类也显得尤为重要。  幸运的是计算机病毒通常都具有家族特性，某个计算机病毒经过一些演化，会繁衍出很多变种病毒。病毒变种的目的主要是针对某些杀毒软件的进行的免杀处理。通常这样的病毒及其繁衍出来的病毒，具有家族特性。如果我们根据病毒的家族特性对病毒进行划分，往往会提取该家族的病毒都存在的特征信息作为病毒的判断依据，从而使得病毒库中一条记录信息就可以命中该家族的全部病毒。由此可见，准确的病毒家族聚类方法可以大幅度提高反病毒软件的查杀效率，减少病毒特征码库的大小。  2．国内外研究现状和发展趋势  由于新的病毒往往不会包括已知的特征代码，导致传统的特征代码法不能检测未知病毒，使检测总是滞后于病毒的爆发。未知病毒检测逐渐成为人们研究网络安全问题的焦点。  虽然 1987 年 cohen的研究表明，无论是推理性还是实时分析，病毒检测都是无法达到绝对的准确。但是人们还是不断用新的技术来达到更好的准确度，新兴的领域如数据挖掘、机器学习和人工智能等给人们带来了新的灵感。  在 IBM，KePhart和 Arnold发明了一种统计的方法来自动提取病毒的特征。该研究基于语音识别算法，经验证该方法在检测已知病毒的过程中可以达到同人类专家几乎一样的效果。Loetal提出了一种恶意代码过滤器，手动地过滤恶意代码的特征。类似的研究还有针对 UNIX 系统的恶意程序特征的过滤器，以及一种半自动检测恶意代码的方法。1996年，IBM的科研人员提出利用人工神经元，这种分类器可成功检测到 80 到 8590种的未知引导扇区病毒，误检率低于 1%，但由于算法本身的局限性，该分类器只能分析引导扇区类的病毒，而这种类型的病毒仅占所有二进制病毒总数的 5%。Gerald、Lee等人把数据挖掘技术应用在入侵检测系统中，他们利用系统调用和网络数据进行机器学习，从而检测新的入侵。  国内在未知病毒检测领域也有很多新的发现，在基于获取PE文件静态特征和 API调用序列提取特征的基础上，各种机器学习和数据挖掘的方法被广泛应用在病毒检测中。四川大学的陈恒、刘晓洁等提出了一种基于人工免疫的病毒检测方法，给出了计算机软件系统中自体、非自体、抗原、免疫细胞等的表示方法，实现了否定选择和克隆选择等免疫机制，能够有效地检测未知计算机病毒。国防科技大学的张波云等提出了基于 K-最近邻算法的未知病毒检测方法、模糊模式识别的方法和基于多重朴素贝叶斯算法的未知病毒检测技术等。  以上各种方法，重点在于实现程序的二分类，即检测出程序是病毒，或者是正常文件，近年来，也有一些研究者开始将他们的注意力放在用不同的方法对病毒的详细分类上。Gheorghescu使用了恶意软件中的基本代码块实现了一个病毒自动分类系统，基本代码块定义是一个不包含跳转或跳转目标的连续的指令序列，两个基本块之间的距离则定义为块和恶意块之间不同的字节个数。Sathyanarayan使用从恶意软件中静态分析提取出 API 来构造整个类的签名，将未经分类的恶意软件的 API 调用序列同已经识别的恶意软件的序列签名进行比较，从而分类出其所属家族。重庆大学的洪群业提出通过提取脱壳后的 PE 文件反汇编包含的字符串信息，以解决检测 PE 文件的病毒分类问题。  3. 关键科技问题及本课题应用前景  在国内外的研究人员的各种检测方法提出了一些卓有成效的检测未知计算机病毒的方法，在利用机器学习、模式识别、数据挖掘等新技术实现病毒检测方面取得了卓有成效的进展。但是也还存在一些有待解决的问题：  （1）使用API调用序列作为特征向量，需要待测样本未经过 API 调用混淆和加密等问题；  （2）对加壳的PE文件脱壳的困难重重，尤其是一些个人制作的未公开的壳，只能是手工进行脱壳；  （3）很多PE病毒文件都进行了加密处理，很难提取出其字符串等信息。  因此，如何根据现有的PE静态文件结构信息，提取出其特征向量对病毒文件进行家族聚类是提高处理速度和效率的关键途径。该课题的研究可以大大减少病毒特征码库中病毒特征码的数量，并且可以提高对新病毒的响应速度和减少其处理时间。  主要参考文献  [1] 韩兰胜, 高昆仑, 赵保华. 基于API函数及其参数相结合的恶意软件行为检测[J]. 计算机应用研究, 2013, (11): 3407-3425.  [2] Zahra Khorsand, Ali Hamzeh. A Novel Compression-Based Approach for Malware Detection Using PE Header[J]. Conference on Information and Knowledge Technology (IKT), 2013, (5): 127-133.  [3] Boris Rozenberg, Ehud Gudes. A Method for Detecting Unknown Malicious Executables[J]. International Joint Conference of IEEE, 2011, (1): 190-196  [4] 洪群业. 基于分类的未知 PE 病毒检测技术的研究 [D]. 重庆大学:洪群业, 2010.  [5] Kolter, J. Zico, and Marcus A. Maloof. Learning to detect and classify malicious executables in the wild. The Journal of Machine Learning Research 7 (2006): 2721-2744.  [6] 段刚. 加密与解密(第三版)[M]. 电子工业出版社, 2012.  [7] Alex Rodriguez, Alessandro Laio. Clustering by fast search and find of density peaks [J]. Science (2014): 1492-1496  [8] 唐东明. 聚类分析及其应用研究[D]. 电子科技大学, 2010.  [9] Ronghua Tian, Lynn Batten, Rafiqul Islam. An Automated Classification System Based on the Strings of Trojan and Virus Families [J]. IEEE, 2009 : 23-30  [10] 王鑫, 姚辉, 刘桂峰. 一种提取pe文件特征的方法及装置[P]. 中国: CN103886229 A,2014年6月25日.  [11] 陈桓,刘晓洁,宋程. 一种基于免疫的计算机病毒检测方法[J]. 计算机应用研究,2005,9:23-26.  [12] Abhishek Karnik, Suchandra Goswami & Ratan Guha. Detecting Obfuscated Viruses Using Cosine Similarity Analysis [J]. IEEE, 2007: 165 – 170. | | | |

|  |
| --- |
| **二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）（不少于2500字）**  现有的病毒聚类划分方法主要采用人工划分的方式进行处理，主要处理方式是通过动态执行病毒文件，人工记录和分析病毒的动态行为特征，比如调用系统应用程序编程接口（API）的顺序，是否修改注册表，是否在敏感位置创建文件等等，之后再结合这些动态的行为特征进行人工聚类划分。对每个未知病毒进行如上处理，势必会浪费大量的人力、财力。  而本课题正是为了弥补人工聚类方法的缺陷：需要大量的人力资源，同时对病毒样本分析人员的经验要求也很高。人工对病毒进行聚类划分时，需要动态运行病毒文件查看其动态行为特征，不但增加了很多时间和计算资源上的消耗，而且动态运行病毒文件对本地计算机也会增加很多染毒风险。  1．课题主要研究内容  （1）分析并研究 PE静态文件结构及信息  PE文件使用的是一个平面地址空间，所有代码和数据都被合并在一起，组成一个很大的结构。文件的内容被分割为不同的区块（Section），区块中包含代码或数据，各个区块按页边界来对齐，区块没有大小限制，是一个连续的结构，每个区块都有它自己在内存中的一套属性，比如：这个块是否包含代码、是否只读或者可读/可写等。  每个PE文件都是以一个DOS程序开始的，一旦程序在DOS下执行，DOS就能识别出这是有效的执行体，然后运行紧随MZ header之后的DOS stub。紧接着 DOS stub 的是 PE header。 PE header 是PE相关结构 IMAGE\_NT\_HEADERS 的简称，其中包含了许多PE装载器用到的重要域。  NT Header中包含着指向IMAGE\_FILE\_HEADER（映像文件头）和IMAGE\_OPTIONAL\_HEADER（可选映像头）的指针。其中映像文件头中包含了文件的区块数目、文件创建日期和时间、文件属性等信息。可选映像头是一个可选结构，但实际上映像头不足以定义PE文件属性，因此可选映像头中定义了更多的数据，其中主要数据有： 说明文件是ROM映像还是普通可执行映像的标志字；所有含代码区块的总大小；程序执行入口的相对虚拟地址（RVA，指向主线程代码。病毒可利用该特性对程序进行重定位，在节表末尾添加一个代码节，然后修改该成员的值指向新节，从而优先程序代码得到执行权，在病毒代码执行完毕后再返回原入口地址即可不被发现）；一个包含16个成员的结构数组（病毒要调用API函数时先通过该成员找到导入表，再进一步得到相关dll中的API函数地址）。  在PE文件头与原始数据之间存在着一个区块表，区块表包含每个块在映像中的信息，分别指向不同的区块实体。每个区块表保存了各个区块的相关信息。其主要信息有：区块名；PE加载在内存后区块数据的RVA（定位代码入口地址等需要根据该值计算相应的文件偏移，来取得入口点代码等）；原始数据的大小；区块属性（包括是否包含可执行代码、初始化数据、对节内存的执行、读写权限等。如果一段非常规命名的节属性为可执行，有可能是被感染文件）。  在区块表之后就是PE文件的原始数据：代码段、数据段、资源段等。  因为PE文件病毒具有家族性，所以每个家族的病毒其PE文件头和区块中必然包含着有别于其他病毒家族的特征。通过从PE文件中提取PE文件头信息、导入表信息、字符串信息和代码段信息并转化为特征向量，以便于接下来用聚类方法对PE病毒文件进行聚类。  （2）研究从PE静态文件中提取结构特征的关键技术  当前，应用最广、最精确和快速的恶意代码检测技术是基于特征码的扫描技术。这种技术的核心是恶意代码特征码库，能够查杀一种或一类恶意代码的前提是在恶意代码特征码库中存有与其相对应的特征码。  对于未加壳恶意代码，一般直接在其程序入口点处选取部分代码作为特征码即可。但是当恶意代码采用加壳和多态技术保护后，一个恶意代码就会出现近似于无穷多种表现形式，使用普通的方法无法正确提取特征码，结果不仅会导致特征码库的爆炸，而且还会出现一个特征码只能查杀一个恶意代码，或使用某种加壳工具加壳的所有程序都被错误查杀的现象。  而绝大多数恶意程序都会采用加壳方式来避开杀毒软件的查杀。这也给从PE静态文件中提取有区分度的结构特征困难重重。而如果尝试对壳进行处理，即先检测该PE文件是否加壳，如果加壳对其脱壳。但是壳的种类很多，已知的、未知的，而我们只能针对特定的壳开发出特定的脱壳软件，而对一些加密壳或者修改壳，没有专门的脱壳机，此时必须要分析外壳，手动脱壳。而这对于本课题的研究目内容来说，大大减低了该系统的效率。所以只能忽略壳，从已有的PE结果中提取出具有区分度的特征值。  （3）分析当前主要的聚类算法，研究适合于对PE文件聚类分析的算法  随着计算机应用的不断推广，现实生活中积累了不断增长的海量数据。数据挖掘方法是解决从海量信息中提取有用知识、提供决策支持的有效途径。数据挖掘的主要功能有分类、聚类、数值预测、时间序列和序列模式挖掘等。分类是其中一个重要方向，就是找出一个类别的概念描述，代表了此类数据的整体信息，并以此构建一个模型，然后用建好的模型就能实现预测类标记未知的对象。  聚类的目标是使同一类对象的相似度尽可能地大；不同类对象之间的相似度尽可能地小。目前聚类的方法很多，根据基本思想的不同，大致可以将聚类算法分为五大类：层次聚类算法、分割聚类算法、基于约束的聚类算法、机器学习中的聚类算法和用于高维度的聚类算法。  1） 典型的基于距离的聚类算法K-means算法  K-means算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。  K个初始类聚类中心点的选取对聚类结果具有较大的影响，因为在该算法第一步中是随机的选取任意K个对象作为初始聚类的中心，初始地代表一个簇。该算法在每次迭代中对数据集中剩余的每个对象，根据其与各个簇中心的距离将每个对象重新赋给最近的簇。当考察完所有数据对象后，一次迭代运算完成，新的聚类中心被计算出来。如果在一次迭代前后，误差平方和的值没有发生变化，说明算法已经收敛。  算法过程如下：  a. 从N个点随机选取K个点作为质心。  b. 对剩余的每个文档测量其到每个质心的距离，并把它归到最近的质心的类  c. 重新计算已经得到的各个类的质心  d. 迭代b～c步直至新的质心与原质心相等或小于指定阈值，算法结束  但是k-means 算法的缺点确定也很明显：  首先，在 K-means 算法中 K 是事先给定的，这个 K 值的选定是非常难以估计的。很多时候，事先并不知道给定的数据集应该分成多少个类别才最合适。这也是 K-means 算法的一个不足。  其次，在 K-means 算法中，首先需要根据初始聚类中心来确定一个初始划分，然后对初始划分进行优化。这个初始聚类中心的选择对聚类结果有较大的影响，一旦初始值选择的不好，可能无法得到有效的聚类结果，这也成为 K-means算法的一个主要问题。  最后，从 K-means 算法框架可以看出，该算法需要不断地进行样本分类调整，不断地计算调整后的新的聚类中心，因此当数据量非常大时，算法的时间开销是非常大的。所以需要对算法的时间复杂度进行分析、改进，提高算法应用范围。  2）快速检索和查找密度峰值的聚类方法  与K-means聚类方法类似，该方法只基于两个数据点之间的距离。它可以检测非球形的数据集，也可以自动的查找出簇的正确的个数。  我们把数据点密度局部最大值作为簇的中心点。这个算法基于这样的假设：我们认为簇的中心点周围是密度比中心点小的局部点，并且这些点与其他密度高的点的相对距离比与簇中心的距离大。对每个数据点，我们计算两个属性：局部密度和与密度比该点大的距离。  算法过程如下：  计算每个点的局部密度。首先选出距离阈值，如果两点距离小于该阈值，那么两点的密度值都加1，否则密度值不变  计算每个点到比其密度高的点中最小的距离d\_min。  将密度最大的点作为簇的一个中心点，选取d\_min与其他点的d\_min相比异常大的点作为簇的中心点。  该算法只对不同点的密度大小敏感，因此对于大数据集来说，阈值的选取影响着聚类结果的鲁棒性。该算法相对于K-means的优点是不用选取簇类个数K值，也不用选取初始聚类中心。而且只需要计算一次密度和距离，不用循环迭代，因此极大的降低了时间复杂度。  2．课题研究目标和效果  目前，反病毒软件的开发严重依赖对病毒样本的分析和检测。分析技术非常重要，但是基本主要依赖分析人员的知识和经验，很多新的病毒其实只是以前代码的变种。如何对大量病毒进行分类，从而将分析人员的精力集中在真正的未知病毒，对于提高病毒分析效率是非常重要的。当前对于病毒检测的研究大部分集中在实时分析病毒特征的技术，因此一个高效的分类系统就成为关键。  本课题的研究目标是完成一个通过对大量的PE病毒文件提取特征向量，经过聚类分析算法对这些病毒进行聚类的系统。通过该系统，将大量的病毒文件根据其特征分为不同的类别，从而使分析人员分析一类病毒的单个文件就可以大致了解该类病毒的特征，为分析人员提取病毒的家族特征码节省大量的精力。  预期完成学术论文1篇，毕业论文1篇。 |

|  |
| --- |
| **三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性等）（不少于800字）**  1．研究方法  针对本课题的研究内容和特性，达到研究目标和完成毕业论文，采用如下的研究方法：  （1）理论知识准备：通过查阅图书馆馆藏书籍、学术资源数据库和技术网站等资源，学习本课题有关理论研究和实际应用，了解其中存在的问题，为本课题的研究做科学的基础知识准备工作。  （2）深入学习研究PE文件的静态信息，结合已有的文献等，提取出有区分度的特征信息。  （3）对课题中涉及到的机器学习、病毒原理等基本理论进行研究和学习，结合本课题的特征做出符合本课题的优化。  （4）设计技术方案：在理论准备和技术调研的基础上，结合需求分析的结果，设计出本课题详细的完整的技术方案。  （5）实现并测试：理论与实践相结合，根据已设计出的技术方案，对需要改进的关键部分进行仿真，验证设计方案的可行性，然后利用已搭建好的环境进行集成测试,  （8）整理文献资料、系统代码和数据等，完成论文。  2．技术路线  （1）分析目前提取PE文件静态信息的原理和技术。Python编程语言具有敏捷开发，方便快捷，而且源代码可读，易懂，作为学习研究非常方便的优点，而且Python中的pefile模块对PE的结构已经做了非常充分的解析，各种关键的数据结构能够非常容易的获得，对于我们做二次开发非常方便。  （2）分析现有的聚类分析算法的原理和实现机制，从中总结找出适合本课题使用研究的聚类算法。由于Python属于解释型语言，处理算法效率比较低，因此选用了Java作为实现算法的平台。  （3）在两个平台之间必然有一个传递信息的方式，计划选用Google的Protocol Buffers作为信息传递中间通道。Protocol Buffers 是一种轻便高效的结构化数据存储格式，可以用于结构化数据串行化，很适合做数据存储或 RPC 数据交换格式。它可用于通讯协议、数据存储等领域的语言无关、平台无关、可扩展的序列化结构数据格式。目前提供了 C++、Java、Python 三种语言的 API。  （4）对模型中提出的关键创新技术进行仿真实现，并分析其优势和特点。  本课题的主要业务流程图如下所示：    3．可行性分析  近年来随着Internet技术的巨大进步，计算机病毒通过网络传播，在对抗反病毒技术上也有很大提高，给网络信息安全带来很大威胁。国内外研究人员在未知病毒检测领域中也取得了不错的进展，其中数据挖掘技术的使用在多数检测方法中起到了重要作用。而且现在国内外有很多成熟的聚类分析算法，对PE文件特征的提取上也有较多的研究这为本课题的研究奠定了坚实的理论背景和强有力的应用支撑。  一方面，Python语言对解析PE文件数据提供了强有力的支持；另一方面Java语言对聚类算法做了较好的支持。选择Google 的Protocol Buffers做两个平台下的通信协议，因为其占用空间小，且python和Java都对其支持。所以，本课题架构和实现上都是切实可行的。  综上所述，本课题是切实可行的。 |

|  |
| --- |
| **四、本研究课题可能的创新之处（不少于500字）**  目前，反病毒软件的开发严重依赖对病毒样本的分析和检测。分析技术非常重要，但是基本主要依赖分析人员的知识和经验，很多新的病毒其实只是以前代码的变种。如何对大量病毒进行分类，从而将分析人员的精力集中在真正的未知病毒，对于提高病毒分析效率是非常重要的。当前对于病毒检测的研究大部分集中在实时分析病毒特征的技术，因此一个高效的分类系统就成为关键。  1、提出了一种提取PE结构信息的方法  目前对PE文件结构信息提取的方法较多，主要有提取其API函数、提取其字符串信息、提取其代码结构特征等。但是对于加壳的PE文件来说，提取其特征，不能只提取某个固定的特征，因为这些特征可能已经在加壳的过程中遭到了破坏，因此这些特征的提取上，存在着很大的困难。  本课题计划将提取尽量多的PE文件结构特征作为特征向量。首先提取PE文件的头信息，比如函数入口点、区块信息表等；其次，提取代码块、资源块、数据块等关键位置的信息；将入口块代码作部分反汇编，去除其跳转位置信息等，保留其进行操作的信息作为对比特征向量。  2、提出将快速检索和寻找密度峰值聚类方法应用于PE文件聚类  相对于经典的聚类算法K-means，密度峰值聚类算法具有如下优势：不需要选取分类个数K，在实际应用中很难预先选出分类个数，尤其是在高维数据中；只需要计算一次，不需要迭代，而K-means算法需要不断地进行样本分类调整，不断地计算调整后的新的聚类中心，因此当数据量非常大时，算法的时间开销是非常大的；不需要选取初始聚类中心点，克服了K-means算法中初始聚类中心点选择不合适影响分类结果。 |
| **五、研究基础与工作条件（1.与本项目相关的研究工作积累基础 2.包括已具备的实验条件，尚缺少的实验条件和拟解决途径）（不少于500字）**  1、研究工作积累基础  本人在实验室担任了IPv4-IPv6业务平台系统的业务管理系统的研发，还有航空公司安全保障能力和SMS审核系统中的Android端逻辑业务的开发，对处理数据和处理逻辑性强的课题积累了大量的经验和教训。并且也一直对计算机病毒和系统漏洞的原理很感兴趣，在此期间，进行了大量的研究和学习。  在实习过程中一直从事于课题相关项目的研究，深入研究了PE文件的结构特征，并对机器学习做了比较深入的研究。这些都为本课题的研究提供了知识储备。  在实习过程中，还采用Python开发网络爬虫，是我掌握了Python开发的基础知识和基本技能，为我顺利使用Python提取PE文件的静态结构特征提供了保证。  在本课题提出之前我阅读研究了大量的有关PE文件特征和聚类算法的文献，分析了该课题的现实需求和实现的可行性，基于此提出了该课题。  所以本课题的提出具有理论依据，同时有实际的项目技术支持。  2、已具备的实验条件  可以熟练的使用Java编程语言和Python编程语言进行开发，并对PE静态病毒文件的特征有一些认识和研究。  对经典的、常用的聚类分析算法有一定的认识和研究，对各种聚类算法的优缺点有一定的认识，且对算法解决特定问题的优劣比较方法有一定研究。  3、缺少的实验条件和拟解决途径  本人虽然对PE文件的主要静态特征进行了提取和分析，但是未得到实践的验证，其是否符合要求的区分度尚未可知。  而且，对聚类算法的本质并没有认识透彻。本人计划更加深入的研究现有的聚类算法和PE文件特征提取点，提出更加具体、科学、可行的理论方案，并进行测试验证。  本课题需要大量的病毒样本进行测试，而病毒样本的获取和保存有比较大的风险，上网找专门的病毒网站，寻找病毒样本，并将其后缀名去掉保存，以避免电脑中毒。 |

**学位论文工作计划**

|  |  |  |
| --- | --- | --- |
| 时间 | 研究内容 | 预期效果 |
| 2014.11-2014.12 | 确定所选题目，了解课题现状及研究背景，撰写开题报告 | 完成开题报告 |
| 2015.01-2015.02 | 研究PE文件特征提取内容和方法 | 确定提取PE文件的特征向量 |
| 2015.03-2015.05 | 查阅和整理分类算法文献材料 | 选择分类算法 |
| 2015.06-2015.08 | 研究系统的实现方法 | 实现本课题预期的系统 |
| 2015.09-2015.11 | 撰写论文 | 完成毕业论文 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 评  定  小  组  成  员 | 姓 名 | 职 称 | 单位名称 | 职务 |
| 孙斌 | 副教授 | 计算机学院 | 成员 |
| 钮心忻 | 教授 | 计算机学院 | 成员 |
| 张茹 | 副教授 | 计算机学院 | 成员 |
| 谷利泽 | 副教授 | 计算机学院 | 成员 |
| 马兆丰 | 讲师 | 计算机学院 | 组长 |
| 导师意见： | | | | |
| 该论文选题“基于密度峰值分类算法的可移植可执行文件病毒聚类分析”，论文对国内研究现状进行了研究分析及对比，在研究反病毒技术的理论基础上，重点研究采用密度峰值分类算法对可移植可执行文件病毒聚类划分，选题具有较强的理论与实际意义。  论文已对所涉及的关键技术进行了较为系统的研究，并在实际的研究与开发中取得了较好的成果，论文对所涉及的“PE静态文件提取结构特征、基于距离的聚类算法K-means算法”作为论文的核心和关键技术进行了深入研究、分析和实验验证，论文具有较好的创新性和可行性。  论文应完成科研成果如下：  1.对论文涉及的相关理论与实现技术进行系统研究与实现，完成1套可完整演示的基于密度峰值分类算法的可移植可执行文件病毒聚类分析系统。  2.对所采用的方法和实现的技术进行功能验证和性能评测，功能要完整实现，性能评测数据不低于6组。  3.要求完成与论文相关的科研小论文1篇并发表见刊。  4.按学校培养要求完成毕业论文1篇并进行查重审核。  同意开题。 | | | | |
| 导师（签名）：  日期： 年 月 日 | | | | |
| 开题报告小组意见： | | | | |
| 组长（签名）：  日期： 年 月 日 | | | | |
| 学院意见（签章）： | | | | |
| 负责人：  日期： 年 月 日 | | | | |