УДК 004.056

Д.В. Комашинский

ПОДХОД К ВЫЯВЛЕНИЮ ВРЕДОНОСНЫХ ДОКУМЕНТОВ НА ОСНОВЕ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Комашинский Д.В. **Подход к выявлению вредоносных документов на основе методов интеллекуального анализа данных.**
Аннотация. Работа посвящена проблеме безопасности файловых объектов формата Portable Document Format. Обобщаются существующие практики, нацеленные на выявление вредоносных документов. Формируется набор основных групп статических признаков вредоносных и безопасных документов. Собранные данные используются для построения системы автоматической классификации новых, ранее неизвестных документов, на основе методов интеллектуального анализа данных (Data Mining). Анализ результатов использования отдельных групп признаков позволяет сформировать новую модель представления документов, основанную на описании взаимосвязей и содержания их основных структурных элементов. Применение полученной модели позволяет оптимизировать целевую функцию систем обнаружения вредоносных документов в базисе требований к точности принятия решения и времени анализа.
Ключевые слова: разрушающие программные воздействия, вредоносные документы, анализ данных, классификация.

Komashinskiy D.V. **An approach to detect malicious documents based on Data Mining techniques.**
Abstract. The research encompasses information security topics related to Portable Document Format. It generalizes existing practices focused on malicious documents detection and forms a set of features which are substantial for deciding whether a document malicious or not. Then the harvested data is adopted for preparing Data Mining - based decision making system which is capable to classify new, previously unknown documents automatically. The obtained accuracy results for distinct feature groups gives an opportunity to design a new representation model for documents. The model is based on static description of main structural elements of documents and their dependencies. The model's usage provides a way to optimize objective function of malicious document detection systems in a requirements basis covering decision accuracy and time.
Keywords: malware, malicious documents, data analysis, classification.

1. Введение. Взрывообразный рост популярности Интернет обозначил ряд новых возможностей как для пользователей сети, так и для злоумышленников. Так, за последнее десятилетие перечень подходов к распространению вредоносных программ (ВП) значительно расширился. Наряду с традиционными схемами внедрения ВП на атакуемые хосты посредством их установки и запуска, важную роль в современном компьютерном мире играет так называемая эксплуатация уязвимостей программных приложений. К примеру, для пользователя персонального компьютера это означает то,

что запуск вредоносной программы на его стороне может быть незаметно осуществлен просто при просмотре им содержимого скомпрометированной Web страницы.

Противостояние данному типу угроз включает разнообразные организационно-технические мероприятия, направленные на разработку систем оценки репутации и категоризации объектов Интернет, улучшение средств проактивного контроля потенциально уязвимых приложений (Host Intrusion Prevention Systems, HIPS) и средств анализа потенциально опасных файлов, обработка которых на стороне пользователя приводит к началу выполнения вредоносного кода. Представленная работа посвящена анализу проблемы выявления вредоносных файлов формата Portable Document Format [13], активно используемого современными средствами эксплуатации уязвимостей.

2. Обнаружение вредоносных документов. Одной из первых публикаций, посвященных комплексному анализу проблемы безопасности электронных документов формата PDF, является работа Блонса и др. [5]. В ней авторы рассматривают среду просмотра и создания документов как среду программирования. Проводится ретроспективный анализ известных на то время потенциальных уязвимостей формата и среды, и рассматриваются возможные сценарии ее злонамеренного использования. Практический анализ статических и динамических особенностей вредоносных документов в дальнейшем проводится Кубеком и др. [7], Ласковым и др. [8], Рахманом [10], Тзермиасом и др. [11] в рамках исследования формата PDF. Отдельно следует отметить работы, посвященные поиску и выявлению вредоносных документов формата OLE, используемых программным пакетом Microsoft Office. В работе Эдвардса и др. [6] проблема анализа документов OLE рассматривается как задача выявления структурных аномалий в отдельных структурах анализируемых файлов. Столфо и Ли [9] представили комбинированный подход к обнаружению документов данного формата как статическими, так и динамическими средствами. Особое внимание в этой работе уделено его структурным особенностям, рассматриваемым в контексте типов угроз, свойственных таким документам.

Анализ результатов, представленных рядом указанных выше работ [7, 10, 11] позволил определить набор групп статических признаков, используемых экспертами для ручного и автоматизированного анализа новых, ранее неизвестных документов. Это позволило автору провести ряд предварительных исследований [1,

4] применимости методов интеллектуального анализа данных [2] (Data Mining, DM) для построения систем автоматического определения степени опасности ранее неизвестных документов формата PDF. Проведенные несколько позднее независимые исследования Ху, Ванга и др. [12] показали непротиворечивость результатов указанных результатов. В качестве используемых групп признаков были выбраны как ранее упомянутые, так и новые группы происхождения документов и типов их информационного наполнения:

1) группа происхождения документов. Применительно к формату PDF (далее это справедливо по отношению ко всем последующим группам признаков), данная группа включает в себя численные и булевы признаки наличия в документе тех или иных информационных атрибутов (пункт 14.3.3 [13]) информационного каталога документа;
2) группа типов информационного наполнения. Включает численные и булевы признаки наличия в документе косвенных объектов тех или иных типов;
3) группа используемых методов компрессии косвенных объектов. Включает численные и булевы признаки использования в документе методов упаковки содержимого потоковых косвенных объектов;
4) группа использования техники обфускации имен. Эта группа охватывает все численные и булевы признаки наличия в документе обфусцированных (пункт 7.3.5 [13]) имен;
5) группа использования автоматических событий, состоящая из численных и булевых признаков наличия в документах основных типов автоматических событий (пункт 12.6.4.1 [13]);
6) группа использования XML форм (XML Forms Architecture, XFA). Включает численные и булевы признаки наличия аномалий в XFA-содержимом;
7) основная группа аномалий и структурных особенностей. Данная группа включает обобщенные данные по признакам предыдущих групп и численные и булевы признаки наличия отклонений от спецификации PDF в основных структурах документа.

Исследование применимости данных групп признаков показало значимость данных об информационном наполнении документов наряду с традиционными группами. Это повлекло за собой проведение

дополнительной обработки доступных данных [3] и формулирование на основе ее результатов излагаемого ниже подхода к выявлению вредоносных документов.

3. Предлагаемый подход.

Статической моделью документа является набор M_{PDF}, включающий ориентированный граф представления документа G и множества типов связей S, типов узлов N и идентификаторов узлов I:$M_{PDF} = (G, S, N, I)$.

Граф представления документа $G = (V, E)$, где V - множество вершин, а E - множество ребер.

Множество типов связей $S = \{s_1, \dots, s_k\}$, где $k = |S|$ – мощность множества связей, и $s_i \in S, 1 \le i \le k$.

Множество типов узлов $N = \{n_1, \dots, n_t\}$, где $t = |N|$ – мощность множества типов узлов, и $n_i \in S, 1 \le i \le t$.

Множество идентификаторов узлов $I = \{i_1, \dots, i_m\}$, где $m = |I|$ – мощность множества идентификаторов узлов, и $i_j \in I, 1 \le j \le m$.

Элементы множества вершин $V = \{v_1, \dots, v_z\}$, представлены как наборы $v_j = \langle i, n \rangle$, где $i \in I, n \in N$.

Элементы множества ребер $E = \{e_1, \dots, e_x\}$, представлены как наборы $e_k = \langle \{v_i, v_j\}, s \rangle$, где $s \in S$ а $v_i, v_j \in V$.

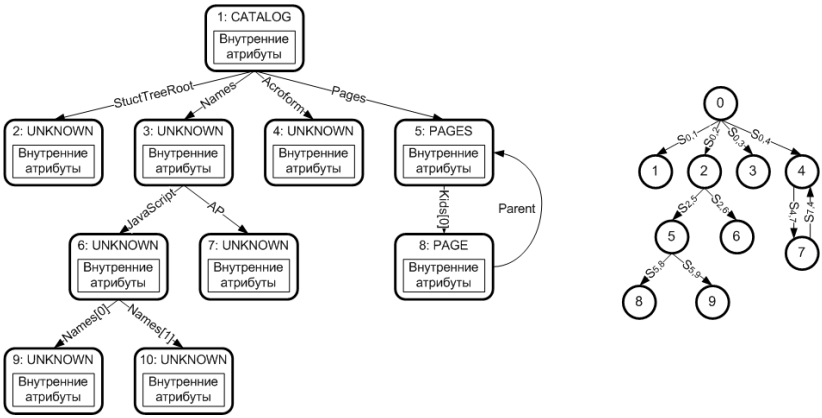

Рис. 1. Пример графа представления документа

На левой части рисунка 1 изображено графическое представление документа. Оно включает набор косвенных объектов (indirect objects) PDF [13] в виде прямоугольников со скругленными

краями и связей между ними, обозначенных стрелками. В верхней части прямоугольников, описывающих косвенные объекты, указаны внутренний номер объекта и его тип.

Например, основной косвенный объект документа «CATALOG» (№1) имеет ссылку типа «Pages» на объект «PAGES» (№5). В соответствии со спецификацией PDF [13], данная связь определяет хранилище каталога всех страниц, содержащихся в документе. Как видно из наличия единственной ссылки «Kids[0]» между объектом «PAGES» (№5) и объектом «PAGE» (№8), документ содержит только одну страницу.

На правой части рисунка 1 представлена упрощенная схема документа для наглядного описания алгоритма извлечения признаков из определенной модели представления.

Таблица 1. **Алгоритм извлечения признаков из файлового объекта PDF**

```
1   Procedure pdf_extract_features(file)
      // file: очередной файловый объект
      // r: индекс вершины графа, сопоставленной с корневым элементом дерева
2     G, r = pdf_parse(file) // получение графа представления G
3     T = pdf_dijkstra(G,r) // получение полного дерева кратчайшего пути T
4     L = pdf_get_leaves(T) // получение списка концевых вершин дерева L
5     R = pdf_new_features_list() // выделение списка признаков R
6     for each l in L do: // перебор списка концевых вершин дерева
7       P = pdf_get_path(T, r, l) // получение пути от корневого элемента до вершины
8       prev_idx = 0 // временная переменная
9       feature = pdf_new_feature() // выделение нового признака
10      for each node_idx in P do // перебор элементов пути
11        if prev_idx then // первый элемент был обработан
12          feature.add(pdf_get_edge(G, prev_idx,node_idx)) // добавление связи
13        end if
14      feature.add(pdf_get_vertex(G, node_idx)) // добавление вершины
15      prev_idx = node_idx
16      end for
17      R.add(feature) // сохранение признака
18    end for
19    return R // возврат списка признаков
```

Для построения набора признаков, представляющих собой цепочки описаний объектов от корневого до конечных элементов иерархии документа, применяется алгоритм Дейкстры (строка 3 таблицы 1). С его помощью получается полное дерево кратчайшего

пути относительно вершины, соответствующей корневому элементу иерархии (строка 4 таблицы 1).

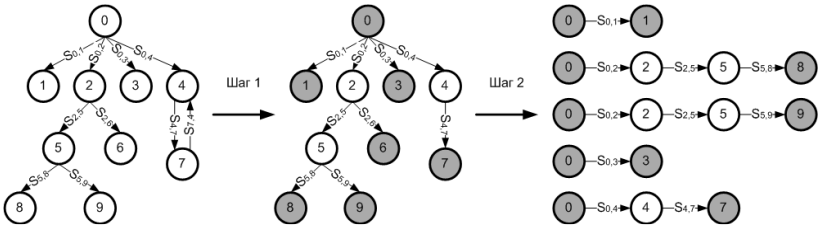

Рис. 2. Процесс извлечения признаков из графа представления документа

Из полученного дерева извлекаются корцевые вершины, используемые для выделения путей в дереве, характеризующих цепочки (признаки), что обозначено в строках 6-19 таблицы 1. Рисунок 2 раскрывает суть проводимых преобразований.

4. Архитектура системы. Для обеспечения вычислительной поддержки исследований используется программный пакет RapidMiner 5.3. На его основе подготовлен базовый набор схем экспериментов, соответствующих канонической форме организации фазы обучения и эксплуатации системы обнаружения вредоносных документов. В качестве основных методов обучения решающих моделей берутся методы Naive Bayes, Decision Tree (C4.5), k Nearest Neighbors. При работе с методами комбинирования решающих моделей используются методы Voting и Stacking. Для оценивания точности получаемых применяется метод десятикратной кросс-валидации. Подпроцесс извлечения признаков для статической методики обнаружения вредоносных документов основан на применении программного средства разбора файлов Open PDF Analysis Framework. Данное средство производит идентификацию и выделение основных структурных частей документ, поиск косвенных объектов и рекурсивный разбор их внутренних структурных элементов.

5. Оценка эффективности. Практические работы по исследованию эффективности предложенной методики состоят из двух основных частей. Первая часть эксперимента посвящена анализу применимости известных структурных особенностей и аномалий вредоносных документов в контексте построения системы их обнаружения на основе DM. Показано, что при использовании отдельных групп структурных признаков, присущих вредоносным документам в соответствии с наблюдениями экспертного сообщества,

общая точность (accuracy) ответов формируемых систем обнаружения вредоносных документов может достигать значения 0.94. При использовании признаков, характеризующих характер информационного наполнения документов можно достичь еще большей точности, близкой к значению 0.99. Последний вывод обуславливает необходимость взгляда на проблему выявления вредоносных документов под другим углом в контексте описанной статической методики обнаружения вредоносных документов (вторая часть практической работы). Эксперименты с ней показали наличие небольшого числа уникальных структурных паттернов, свойственных более чем в 85% обработанных вредоносных документов при полном отсутствии ложных срабатываний. В отличие от существующих статических подходов на основе выявления структурных аномалий, предлагаемый подход ориентирован на обобщение знаний об особенностях внутренней организации вредоносных документов в контексте содержимого их составных элементов и связей между ними. Процедура принятия решений на его основе не требует существенных временных затрат, присущих полному циклу анализа документов PDF, обусловленных необходимостью глубокого структурного разбора документов и распаковки его содержимого, отображения его структурных свойств на общую объектную модель документов PDF и ее динамическую интерпретацию.

6. Заключение. В работе предложена статическая модель представления документов формата Portable Document Format и разработана методика идентификации вредоносных документов на ее основе. Показан факт наличия ограниченного количества структурных паттернов, используемых во вредоносных документах. Использование предложенной методики позволяет строить комплексные системы раннего обнаружения Internet угроз, эксплуатирующих уязвимости приложений.

Литература

1. *Комашинский Д.В., Котенко И.В.* Исследование структурных особенностей вредоносных документов методами Data Mining // Информационные технологии и вычислительные системы, №2, 2012. С.76-92.
2. Кома*шинский Д.В., Котенко И.В.* Концептуальные основы использования методов Data Mining для обнаружения вредоносного программного обеспечения // Защита информации. Инсайд, 2010. № 2, С.74-82.
3. *Комашинский Д.В., Котенко И.В.* Метод извлечения структурных признаков для задачи обнаружения вредоносного программного обеспечения // Изв. вузов. Приборостроение, Т.55, № 11, 2012, С.58-62.

4. *Комашинский Д.В., Котенко И.В.* Обнаружение вредоносных документов формата PDF на основе интеллектуального анализа данных. // Проблемы информационной безопасности. Компьютерные системы. №1, 2012. С. 19-35.
5. *Blonce A., Filiol E., Frayssignes L.* Portable Document Format (PDF) Security Analysis and Malware Threats // Presentations of Europe BlackHat 2008 Conference, 2008.
6. *Edwards S., Baccas P.* Fast Fingerprinting of OLE2 Files: Heuristics for Detection of Exploited OLE2 Files based on Specification Non-conformance // Proceeding of Virus Bulletin Annual Conference, Barcelona, October 2011. P.172-185
7. *Kubec J., Sejtko J.* X Is Not Enough! Grab the PDF by the Tail! // Proceeding of Virus Bulletin Annual Conference, Barselona, October 2011. P.128-135.
8. *Laskov P., Srndic N.* Static Detection of Malicious JavaScript-Bearing PDF Documents // Proceedings ACSAC'11 Proceedings of the 27th Annual Computer Security Applications Conference, 2010. P.373-382.
9. *Li W.-J., Stolfo S.* SPARSE: A Hybrid System to Detect Malcode-Bearing Documents CU Tech. Report, Jan 2008 https://mice.cs.columbia.edu/getTechreport.php?techreportID=504, посещен 23.03.2013.
10. *Rahman M.* Getting Owned By Malicious PDF - Analysis // SANS Institute Reading Room Site, http://www.sans.org/reading_room/whitepapers/malicious/owned-malicious-pdf-analysis_33443 , 2010, посещен 23.03.2013.
11. *Tzermias Z., Sykiotakis G., Polychronakis M., Markatos E.* Combining Static and Dynamic Analysis for the Detection of Malicious Documents // Proceedings of the Fourth European Workshop on System Security, ACM New York, 2011.
12. *Xu W., Wang X., Zhang Y., Xie H.* A Fast and Precise Malicious PDF Filter // Proceedings of 22nd Virus Bulletin Conference, 2012, P. 14-19.
13. International Organization for Standardization, Portable Document Format, ISO 32000-1:2008, http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/PDF32000_2008.pdf, 2008, посещен 23.03.2013.

Комашинский Дмитрий Владимирович – аспирант лаборатории Проблем компьютерной безопасности СПИИРАН. Область научных интересов: обнаружение и анализ вредоносных программ, машинное обучение. Число научных публикаций ― 27. komashinskiy@comsec.spb.ru, http://comsec.spb.ru/ru/staff/komashinskiy; СПИИРАН, 14 линия, 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-2642, факс +7(812)328-4450. Научный руководитель — И.В. Котенко.

Komashinskiy Dmitriy Vladimirovich — postgraduate student, Laboratory of Computer Security Problems, SPIIRAS. Research interests: intrusion detection and analysis, machine learning. The number of publications — 27. komashinskiy@comsec.spb.ru, http://comsec.spb.ru/ru/staff/komashinskiy; SPIIRAS, 14-th line, 39, St. Petersburg, 199178, Russia; office phone +7(812) 328–2642, fax +7(812)328–4450. Scientific advisor — I.V. Kotenko.

Поддержка исследований. В публикации представлены результаты исследований, поддержанные Министерством образования и науки Российской Федерации (государственный контракт 11.519.11.4008), грантами РФФИ, программой фундаментальных исследований ОНИТ РАН и проектами Седьмой рамочной программы Европейского Союза SecFutur и MASSIF.

Рекомендовано лабораторией Проблем компьютерной безопасности СПИИРАН, заведующий лабораторией Котенко И.В., д-р техн. наук, проф.
Статья поступила в редакцию 23.03.2013.

РЕФЕРАТ

Комашинский Д.В. **Подход к выявлению вредоносных документов на основе методов интеллектуального анализа данных.**

Распространение ряда современных Internet угроз осуществляется с помощью так называемых программных пакетов использования уязвимостей (exploit kits). Для достижения поставленной цели они осуществляют сложные многоходовые комбинации, использующие элементы социальной инженерии и программные средства (1) идентификации уязвимых программных приложений атакуемого хоста, (2) автоматического создания средств начального внедрения на него, (3) обеспечения скрытности функционирования установленных вредоносных программ, их обновления и т.д. Данное исследование посвящено анализу проблемы выявления вредоносных файлов формата Portable Document Format, используемых злоумышленниками на этапе начального проникновения вредоносных программ на атакуемый объект. Ее решение в рамках комплексных систем противодействия вредоносным программам позволяет своевременно идентифицировать опасные Web сервера и тем самым минимизировать наносимый пользователям ущерб.

Начальной мотивацией исследования являлась необходимость проверки применимости хорошо зарекомендовавших себя в других областях информационной безопасности методов интеллектуального анализа данных (Data Mining) для построения систем автоматического выявления новых, ранее неизвестных вредоносных документов. Проведено обобщение существующих подходов к их выявлению и сформирован перечень основных признаков вредоносности. Полученные наборы признаков использованы для проведения обучения системы и оценки точности ее решений. Результаты проведенных экспериментов показали, что наряду с традиционно эффективными группами признаков (используемых для выявления вложенного программного кода, фактов упаковки и статической обфускации критических элементов документов), важную роль в процессе принятия решения играют некоторые семантически нейтральные признаки, характеризующие структурные особенности документов и характер их содержимого.

Полученные результаты позволили разработать модель представления документов, основанную на описании содержимого и взаимосвязей их основных структурных элементов. На ее основе предложен новый подход к построению автоматической системы выявления опасных документов. В отличие от существующих, он опирается на обобщение знаний об особенностях программной реализации средств их создания, что обеспечивает общее упрощение процедуры анализа при повышении показателей точности принятия решения.

Архитектура системы включает программные средства статического разбора документов формата Portable Document Format и выделения значимых для процесса принятия решения признаков. Обработка данных осуществляется с помощью программного пакета Rapid Miner 5.3.

SUMMARY

Komashinskiy D.V. **An approach to detection malicious documents based on Data Mining techniques.**

Processes of propagating some modern Internet threats are supported by means of so called Web malware exploitation kits. To this end they implement complex multistep combinations using social engineering and software packages focused on (1) identification vulnerable software applications installed on attacked hosts, (2) automatic creation of malicious files for initial infiltration into attacked hosts, (3) providing secretive and stable functioning of installed malicious software and its prompt updating and so forth. The research is dedicated to analyzing problems of detecting malicious Portable Document Format files which are usually used by malefactors at initial steps of malware infiltration into attacked objects. Their solving in the context of modern multilayer malware counteraction systems gives opportunities to identify and block potentially dangerous Web servers in time and therefore provides the chance to mitigate potential damages.

The basic motivation of the work is a necessity to check whether admittedly efficient Data Mining - based approaches on malware detection are applicable for building automatic systems responsible for detecting new, previously unseen documents. Existing ways of detecting such objects were generalized. Sets of potentially useful and informative static features were prepared. They were adopted in a processes of learning new classification schemes and their validating. Obtained experimental results showed that some semantically neutral features characterizing structural and content traits can be as efficient as well known attributes describing embedded JavaScript code, compression and obfuscation methods and so on.

Thus the results gave an opportunity to develop a new document representation model which is based on describing content and relations of documents' main structural elements. New Data Mining - based approach to construct automatic systems for detecting malicious documents is proposed. In contrast to existing ones, it focuses on generalizing knowledge about traits of means used for automatic creation of malicious documents what makes file analysis steps easy and in many cases improves accuracy of decision making.

The architecture of the implemented system's prototype uses static means of parsing Portable Document Format files and extracting used features. Data processing is performed by RapidMiner 5.3 software package.