УДК 004.02

М.И. Поддубный
(Краснодарское высшее военное училище имени генерала армии С.М. Штеменко;
e-mail: podd.maxim@yandex.ru)

МЕТОДИКА АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ СУЩНОСТЕЙ СИСТЕМЫ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА ПО МЕТКАМ КОНФИДЕНЦИАЛЬНОСТИ

Представлена методика автоматической классификации сущностей системы электронного документооборота по меткам конфиденциальности. Она реализована на основе теории конечных предикатов, которая позволяет, в отличие от ранее известных, классифицировать как пассивные сущности (объекты), так и активные сущности (субъекты) по меткам конфиденциальности.

Ключевые слова: безопасность информации, метка конфиденциальности, автоматическая классификация.

M.I. Poddubny

AUTOMATIC CLASSIFICATION'S METHOD OF ELECTRONIC DOCUMENT MANAGEMENT SYSTEM ENTITIES BY CONFIDENTIAL MARKS

Automatic classification's method of electronic document circulation's entities by confidential marks is represented. It is realized on the basis of finite predicates' theory, which allows classifying passive entities (objects), as well as active entities (subjects), by confidential marks, in contradistinction to the methods, which were developed earlier.

Key words: information safety, mark of confidentiality, automatic classification, predicate.

Статья поступила в редакцию Интернет-журнала 21 ноября 2016 г.

Введение

Стратегия создания в России единой системы информационно-телекоммуникационного обеспечения в интересах государственного управления, обороны страны, национальной безопасности и правопорядка получила развитие, в том числе, в виде ***системы электронного документооборота (СЭД)***.

В зависимости от обрабатываемой информации, СЭД могут использовать различные модели безопасности, в том числе – разнородные, при которых неизбежно возникает необходимость определения и присвоения меток конфиденциальности всем элементам модели, реализующимся, как правило, "вручную". Мало того, что такой подход трудоёмок, он слабо обеспечен методически. Для облегчения и ускорения указанных процессов необходимо автоматизировать процедуру классификации сущностей СЭД по меткам конфиденциальности.

Методические вопросы автоматической классификации сущностей в системе электронного документооборота

Сущности СЭД подразделяются на пассивные (объекты) и активные (субъекты) [1].

Автоматизировать присвоение меток конфиденциальности объектам предлагается способом автоматической классификации конфиденциальных документов в системе электронного документооборота [2].

Присвоение меток конфиденциальности субъектам СЭД, с точки зрения моделей безопасности [3], возможно без участия администратора в двух случаях: по заранее сформированному реестру, строго указывающему каждому пользователю его уровень доступа, и на основе признаков, присущих субъекту системы и позволяющих сделать вывод о его уровне доступа. При взаимодействии локальных сегментов СЭД набор классификационных признаков субъектов системы предлагается формировать, исходя из разрешений дискреционной модели. В качестве разнородных моделей безопасности рассмотрим модель Белла-ЛаПадула (БЛП) и классическую модель Харрисона-Руззо-Ульмана (ХРУ) [4].

Описание методики целесообразно разделить на два этапа, отражающих определение меток конфиденциальности объектов, с одной стороны, и субъектов – с другой.

На ***первом этапе*** для определения меток конфиденциальности объектов использован способ автоматической классификации конфиденциальных формализованных документов в системе электронного документооборота, реализованный в работе [2]. Адекватность этого способа доказывается в работе [5]. При использовании указанного способа каждый объект системы рассматривается как документ, обладающий формальной и информативной частями.

Способ характеризуется следующими элементами [2]:

$O = \{o_j\}$ – множество объектов (документов), поступающих в автоматизированную систему, где $j = \overline{1, j'}$, j' – количество объектов системы;

$Z = \{z_d\}$ – множество реквизитов объекта, где $d = \overline{1, d'}$, d' – количество применяемых реквизитов;

$L = \{l_g\}$ – множество ключевых слов, где $g = \overline{1, g'}$, g' – количество применяемых ключевых слов;

$T = \{t_h\}$ – множество характеристик текста, где $h = \overline{1, h'}$, h' – количество используемых характеристик текста;

$V = \{v_k\}$ – множество форм объекта, где $k = \overline{1, k'}$, k' – количество применяемых форм объекта;

$W = \{w_e\}$ – множество значимых слов в тексте объекта определённой информационной области, обладающих определённым весом, где $e = \overline{1, e'}$, e' – количество применяемых значимых слов;

$U = \{u_p\}$ – множество областей информационной ответственности, определённых в системе, где $p = \overline{1, p'}$, p' – количество областей информационной ответственности;

$(M, \leq)$ – решётка меток конфиденциальности, определённых в системе, $M = \{m\}$, где $m = \overline{1, m'}$, m' – количество видов меток конфиденциальности в системе. В качестве примера в статье рассматривается $m' = 4$, например: $M = \{1, 2, 3, 4\}$, где $1 < 2 < 3 < 4$;

$M_o = \{m_o\}$ – множество меток конфиденциальности объектов, где $o \in O$ – объекты системы.

Кроме того, необходимо учитывать алгоритм извлечения значимых слов и их взвешивания (возможно применение любого из известных алгоритмов, например – Портера) [6], а также предикаты распознавания: реквизитов объекта $P_Z(T, L)$, формы объекта $P_V(Z, L)$, областей информационной ответственности $P_U(W)$, меток конфиденциальности объектов $P_{M_o}(U, Z)$, а также правила их построения.

Этапы реализации способа состоят в следующем [2]:

- в зависимости от признаков документа t_h и ключевых слов l_g, применяемых в данном признаке, с применением предиката $P_Z(T, L)$ определяются реквизиты объекта z_d;
- по ключевым словам l_g и реквизитам z_d, с использованием предиката $P_V(Z, L)$, определяется форма v_k поступившего объекта;
- в зависимости от формы объекта выбираются заранее определённые информационные области, в которых производится анализ значимых слов и их взвешивание [6];
- веса значимых слов w_e, полученные в информационных областях, сравниваются с определенными на этапе обучения значениями, после чего, с использованием системы предикатов $P_U(W)$, определяется область информационной ответственности u_p;
- по областям информационной ответственности u_p и реквизитам объекта z_d, на основе системы предикатов $P_{M_o}(U, Z)$, определяется метка конфиденциальности объекта.

Согласно предлагаемому способу, каждый объект o_j в рамках СЭД обладает определённым набором признаков, необходимых для его классификации, анализируемых на этапе обучения системы и сохраняемых в виде систем предикатов. Таким образом, классификация возможна на любом этапе обработки объекта.

На ***втором этапе*** осуществляется классификация субъектов, которые не обладают, подобно объектам, набором "собственных" признаков и все отличия приобретают в зависимости от своих прав относительно других сущностей системы. При этом определены:

- начальное состояние дискреционной системы;
- условия, позволяющие сделать вывод о безопасности классификации;
- правило построения системы предикатов распознавания меток конфиденциальности субъектов СЭД.

Дополнительно введём обозначения:

$S = \{s_i\}$ – множество субъектов системы, где $i = \overline{1, i'}$, i' – количество субъектов, $S \subseteq O$;

$R = \{r_y\}$ – множество видов прав доступа субъектов к объектам, где $y = \overline{1, y'}$, y' – количество применяемых в системе видов прав доступа, например, $\{read, write, own\}$;

$MT[s, o] \subseteq R$ – матрица прав доступа, строки которой соответствуют субъектам, а столбцы – объектам;

$M_s = \{m_s\}$ – множество меток конфиденциальности субъектов системы S;

предикаты распознавания условий безопасности классификации субъектов P_{X_s} и меток конфиденциальности субъектов $P_{M_s}(M_o, M_{\text{ТЕК}_s}, X_s)$.

Как было указано, методика классификации позиционируется с точки зрения взаимодействия разнородных моделей, то есть существует некоторое состояние дискреционной модели q_0, в котором каждому субъекту соответствует строка матрицы прав доступа, ограничивающая его "возможности" [3]. Исходя из этого условия, логично за начальное состояние системы принять $q_0(S, O, MT)$, где СЭД, субъекты которой будут классифицироваться по уровням доступа, не обладает такими элементами, как метки конфиденциальности.

Начальное состояние q_0 должно быть безопасным. Допустим, что начальное состояние системы q_0 небезопасно, тогда, независимо от того, как и в какое состояние будет переходить система, нельзя будет утверждать о безопасности обрабатываемой информации. Однако доказательство безопасности исходного состояния модели требует отдельного исследования и в рамках данной статьи его результаты не рассматриваются.

В состоянии q_0 (в рамках модели ХРУ) каждому классифицируемому субъекту свойственны [3]:

- множество объектов, к которому имеется право доступа – *read* (чтение);
- множество объектов, к которому имеется право доступа – *write* (запись);
- множество объектов, к которому применено право – *own* (владение).

Так как начальное состояние системы безопасно, то классифицировать субъект предлагается, исходя из разрешений матрицы доступа MT в q_0, меток конфиденциальности объектов, определяемых первым этапом классификации и мандатной моделью, с которой осуществляется взаимодействие [2].

Для каждого классифицируемого субъекта системы формируется множество анализируемых объектов, доступ к которым расценивается системой как основание для присвоения соответствующей метки конфиденциальности. Для обеспечения невозможности получения субъектом завышенной метки конфиденциальности введём понятие условий безопасной классификации X_s. Данные условия требуют отдельного исследования и в статье не рассматриваются.

С целью разработки правила построения системы предикатов (используемых для распознавания метки конфиденциальности субъектов СЭД) введём множество переменных m_o, m_s, x_s с величиной алфавитов 4, 4, 2 соответственно, где m_o – метка конфиденциальности анализируемого объекта; m_s – текущая метка конфиденциальности классифицируемого субъекта; x_s – соблюдение условий безопасности классификации субъекта модели.

Структуру используемых в данном примере признаков представим в виде табл. 1.

Таблица 1

Пример определения метки конфиденциальности для $m' = 4$

Применяемые признаки / Метка конфиденциальности	Метка конфиденциальности анализируемого объекта m_o				Текущая метка конфиденциальности классифицируемого субъекта m_s				Соблюдение условий без-й классификации x_s		Предикат, описывающий процесс классификации субъекта по метке конфиденциальности $P_{M_s}(M_o, M_{\text{ТЕК}_s}, X_s)$
	1	2	3	4	1	2	3	4	1	2	
Метка 4				+	+	+	+		+		$m_o^4 \overline{m_s^4} x_s^1$
Метка 3			+		+	+			+		$m_o^3 (\overline{m_s^3 \vee m_s^4}) x_s^1$
Метка 2		+			+				+		$m_o^2 \overline{(m_s^2 \vee m_s^3 \vee m_s^4)}\, x_s^1$
Метка 1	+								+		$m_o^1 \overline{(m_s^1 \vee m_s^2 \vee m_s^3 \vee m_s^4)} x_s^1$
Сохранение сост. 4	+	+	+	+				+	+		$(m_o^1 \vee m_o^2 \vee m_o^3 \vee m_o^4) m_s^4 x_s^1$
Сохранение сост. 3	+	+	+				+		+		$(m_o^1 \vee m_o^2 \vee m_o^3) m_s^3 x_s^1$
Сохранение сост. 2	+	+				+			+		$(m_o^1 \vee m_o^2) m_s^2 x_s^1$
Сохранение сост. 1	+				+				+		$m_o^1 m_s^1 x_s^1$
Сохранение текущего состояния	+	+	+	+	+	+	+	+		+	$(m_o^1 \vee m_o^2 \vee m_o^3 \vee m_o^4) \cdot (m_s^1 \vee m_s^2 \vee m_s^3 m_s^4) x_s^2$

где $P(M_o)$ – предикат узнавания метки конфиденциальности анализируемого объекта;
$P(M_{\text{тек}_s})$ – текущая метка конфиденциальности классифицируемого субъекта;
$P(X_s)$ – предикат узнавания соблюдения условий безопасной классификации

Однозначность и правильность классификации субъекта для $m' = 4$ меток конфиденциальности доказаны прямым перебором. Это можно проделать для любой m'. При этом для реакции системы на присвоение метки проверяется не наличие необходимых условий, а отсутствие препятствующих, так как при анализе первого документа в потоке у субъекта отсутствует текущая метка классифицируемого субъекта. Включение в табл. 1 предикатов узнавания "сохранения состояний" необходимо, так как система должна осуществлять классификацию всех субъектов системы, в том числе, имеющих права доступа к потенциально опасным объектам, не допуская излишних срабатываний.

Применяя алгебру конечных предикатов [7], составлена система предикатов определения метки конфиденциальности субъекта:

$$\begin{cases} \text{метка } 4 = \left(m_o^4 \overline{m_s^4} \vee (m_o^1 \vee m_o^2 \vee m_o^3 \vee m_o^4) m_s^4\right) x_s^1; \\ \text{метка } 3 = \left(m_o^3 \left(\overline{m_s^3 \vee m_s^4}\right) \vee (m_o^1 \vee m_o^2 \vee m_o^3) m_s^3\right) x_s^1; \\ \text{метка } 2 = \left(m_o^2 \left(\overline{m_s^2 \vee m_s^3 \vee m_s^4}\right) \vee (m_o^1 \vee m_o^2) m_s^2\right) x_s^1; \\ \text{метка } 1 = \left(m_o^1 \left(\overline{m_s^1 \vee m_s^2 \vee m_s^3 \vee m_s^4}\right) \vee m_o^1 m_s^1\right) x_s^1; \\ \text{текущее состояние} = (m_o^1 \vee m_o^2 \vee m_o^3 \vee m_o^4)(m_s^1 \vee m_s^2 \vee m_s^3 \vee m_s^4) x_s^2. \end{cases}$$

Таким образом, правило построения системы предикатов (для узнавания метки конфиденциальности субъектов) можно выразить формулой:

$$P\left(M_s, M_o, M_{\text{ТЕК}_s}, X_s\right) = \bigwedge\left(\bigvee_{\forall m_o = 1} m_o^n, \bigvee_{\forall m_s = 1} m_s, \overline{\bigvee_{\forall m_s = 0} m_s}, x_s\right),$$

где m_o^n – предикат узнавания значения n метки конфиденциальности m анализируемого объекта o_i;

m_s – предикат узнавания текущей метки конфиденциальности классифицируемого субъекта;

x_s – предикат узнавания соблюдения условий безопасности классификации.

Для инициализации классификатора служит режим обучения. При этом должно быть задано множество обучающих сущностей (как объектов, так и субъектов), классифицированных заранее и вручную по меткам конфиденциальности. После извлечения из объектов текстового содержания происходит построение словаря значимых слов. Затем на основе правил формируются системы предикатов $P_{M_o}(U, Z)$ и $P_{M_s}\left(M_o, M_{\text{ТЕК}_s}, X_s\right)$, количество которых определяется количеством меток конфиденциальности, заданных в автоматизированной системе (по которым необходимо классифицировать сущности). Системы предикатов сохраняются в базе данных.

Выводы

Описанная в статье методика позволяет автоматически классифицировать сущности СЭД по меткам конфиденциальности, которые соответствуют правам доступа субъектов в состоянии q_0. Безопасность применения методики обеспечивается соблюдением условий безопасной классификации.

На основе представленной методики могут строиться классификаторы, обеспечивающие автоматизацию функций субъекта-администратора по определению и присвоению меток конфиденциальности при взаимодействии систем разграничения доступа, реализующих разнородные модели безопасности.

Литература

1. ***Гайдамакин Н.А.*** Теоретические основы компьютерной безопасности: учебное пособие. Екатеринбург: 2008. 212 с.

2. ***Поддубный М.И., Королев И.Д., Малышев Д.В., Шайков И.Н.*** Способ автоматической классификации конфиденциальных формализованных документов в системе электронного документооборота // Телекоммуникации. № 8. М.: МГТУ им. Баумана, 2016. С. 18-22.

3. ***Девянин П.Н.*** Модели безопасности компьютерных систем: учебное пособие. 2-е изд. М.: Горячая линия – Телеком, 2013. 337 с.

4. ***Поддубный М.И., Королев И.Д.*** Представление политики мандатного разграничения доступа через модель Харрисона-Руззо-Ульмана // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. Краснодар: КубГАУ, 2015. № 3 (107). http://ej.kubagro.ru/2015/03/pdf/111.pdf.

5. ***Поддубный М.И., Королев И.Д.*** Анализ безопасности информации при применении модели отнесения документов автоматизированной системы к информационным областям ответственности исполнителей // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. Краснодар: КубГау, 2013. № 9 (093). IDA. http://ej.kubagro.ru/2013/09/pdf/42.pdf.

6. ***Craven M., DiPasquo D, Freitag D.*** Learning to Construct Knowledge Bases from the World Wide Web // Artificial Intelligence. 2000. Vol. 118 (1-2). Pp. 69-113.

7. ***Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П.*** Об алгебре конечных предикатов // Научно-технический журнал "Бионика интеллекта". № 3 (77). 2011. С. 3-13.