Построение системы защиты электронных библиотек от несанкционированного копирования документов*

© Ивашко Е. Е.

Карельский научный центр РАН
Институт прикладных математических исследований
г. Петрозаводск
ivashko@krc.karelia.ru

Аннотация

В работе рассматриваются вопросы применения аномального (статистического) подхода к обнаружению вторжений для построения системы защиты электронных библиотек от несанкционированного копирования документов. Предлагается метод создания классификаторов поведения и формирования шаблонов «нормального» поведения пользователя, основанный на построении Марковской цепи.

1 Введение

Электронные библиотеки (ЭБ), как существующие, так и разрабатываемые в настоящее время, обладают большим разнообразием поддерживаемых в них информационных ресурсов, способов организации коллекций, функциональными возможностями пользовательских интерфейсов, архитектурными и другими техническими и технологическими особенностями [5]. Одной из ключевых характеристик ЭБ является объем и качество электронного фонда. При этом следует отметить, что на создание и обработку цифровых коллекций документов тратятся большие материальные и человеческие ресурсы. В связи с этим при построении и эксплуатации ЭБ являются важными задачи защиты электронной информации, связанные с обеспечением:

1. целостности обрабатываемой, передаваемой и хранимой в ЭБ информации;
2. доступности всех открытых для пользователей электронных информационных ресурсов;
3. конфиденциальности персональных и персонализированных данных пользователей ЭБ.

Как правило, вопросы защиты электронных библиотек на сегодняшний день затрагивают главным образом проблемы разграничения доступа (построение дискреционных или мандатных правил) и обмена информацией с доверенными пользователями [5, 7]. Также остается актуальной проблема соблюдения авторских прав при публикации цифровых документов, для решения которой применяются, в частности, различные технические меры, связанные с ограничением распространения и использования электронных копий цифровых документов [5, 12, 10].

Однако, помимо решения традиционных, и в целом технически отработанных, задач защиты электронной информации становятся актуальными и новые задачи обеспечения информационной безопасности, связанные, например, с проблемой защиты ЭБ от несанкционированного полного копирования документов. При этом под полным копированием здесь понимается получение электронных копий всех или большой части цифровых документов ЭБ без согласия ее владельцев.

Актуальность и практическая значимость этой задачи определяется на наш взгляд тем, что имея полную несанкционированную копию электронного фонда злоумышленник может выдать свой сайт за сайт скопированной ЭБ и своими безответственными и/или неквалифицированными действиями скомпрометировать настоящих создателей электронного ресурса, а также использовать скопированные цифровые документы для получения прибыли в обход интересов правообладателей.

В настоящее время вопросы защиты ЭБ от полного несанкционированного копирования, как правило, решаются следующим образом.

1. Игнорирование.

Создатели ЭБ сознательно не создают препятствий для возможного полного копирования своих электронных фондов. Ограничения накладываются лишь с целью соблюдения авторских прав на цифровой документ (т.е. необходимость указания автора документа и источника его получения) в виде специального

соглашения с пользователем по использованию информационных ресурсов (например, как это практикуется в Aladin Digital Library, Gresham College Archives [10]).

2. Ограничение круга пользователей, имеющих доступ к ресурсам.

Круг пользователей, имеющих доступ к ресурсам ЭБ, сознательно ограничивается и включает лишь сотрудников и аффилированных с ЭБ лиц (например, как это сделано в Dartmouth College Digital Library, Harvard University Library [10]).

Проблема несанкционированного полного копирования документов, несмотря на имеющиеся российские и международные законодательные акты, редко находит свое отражение в виде судебных исков в силу различных технических и юридических сложностей. Однако упоминания об актуальности этой проблемы можно найти в условиях пользования различными электронными библиотеками (например, в [8]).

В представленной работе рассматриваются вопросы применения аномального (статистического) подхода к обнаружению вторжений для построения системы защиты ЭБ от несанкционированного полного копирования цифровых документов. При этом попытка полного копирования считается несанкционированным вторжением в компьютерную систему. Для построения системы защиты используется технология моделирования безопасного с точки зрения системы поведения пользователя, под которым здесь и далее понимается специфическая для каждого пользователя последовательность обращений к отдельным цифровым документам коллекций ЭБ. При этом используется подход к созданию шаблонов «нормального» поведения, основанный на анализе поведения пользователей при доступе к документам ЭБ и построении Марковской цепи [9, 11, 3].

Аномальный подход основан на предположении о том, что вторжение (здесь и далее вторжение и несанкционированное полное копирование документов трактуются в одном и том же смысле) проявляется как отклонение от обычного («нормального») или ожидаемого поведения пользователя, которое может быть обнаружено системой путем сравнения с некоторым заданным «шаблонным» поведением.

При использовании аномального подхода для обнаружения вторжений в компьютерную систему возникают следующие основные проблемы:

- построение профиля (шаблона «нормального» поведения) пользователя. Это трудноформализуемая и времяемкая задача, для решения которой требуется проанализировать большое количество априорных сведений о поведении пользователей, связанных с доступом к отдельным цифровым документам конкретной ЭБ;
- определение граничных значений характеристик поведения пользователя для снижения вероятности появления ошибок классификации;
- периодическое обновление шаблонов «нормального» поведения пользователей, вызываемое изменениями в их «нормальном» поведении.

В данной работе основное внимание уделено решению задачи построения шаблона «нормального» поведения пользователя, связанного с доступом к цифровым документам ЭБ.

2 Постановка задачи. Основные термины и определения

Основная цель создаваемой системы защиты заключается в том, чтобы своевременно обнаружить пользователя ЭБ, осуществляющего несанкционированное полное копирование документов библиотеки. При этом ставится задача распознавать и такие ситуации, когда пользователь маскирует свои противоправные действия путем перемежевания процесса полного копирования с обычной работой с ресурсом.

В основу разрабатываемой системы защиты ЭБ от полного копирования положена идея, суть которой заключается в следующем. На основе данных о поведении пользователей, связанных с доступом к документам ЭБ, автоматически выявляются тематические связи между документами, которые и будут являться основой для создания шаблона «нормального» поведения пользователей. Используя сформированный шаблон, система защиты от копирования будет отслеживать действия пользователей при работе с документами ЭБ в плане обнаружения «аномальности» в его действиях, которая в нашем случае трактуется как многочисленные запросы к документам различной тематики. При превышении количества аномальных действий в одной сессии некоторого заданного порога система может классифицировать действия данного пользователя как вторжение.

Для формирования шаблонов «нормального» поведения необходим тренировочный набор заведомо безопасных для ЭБ действий пользователя. В качестве таких исходных данных рассматриваются данные об осуществленных пользователями переходах между документами ЭБ.

Исходные для построения шаблона данные имеют следующую структуру:

1. файл исходных данных состоит из множества следов;
2. след – это упорядоченное по времени множество записей об атомарных действиях пользователя ЭБ, осуществленных за одну сессию работы пользователя;
3. сессия работы пользователя – временной интервал, заключающийся между событиями

входа (авторизации) пользователя в ЭБ и выхода (осуществления последнего атомарного действия) из нее;

4. атомарное действие пользователя – обращение (загрузка, копирование) пользователя к документу ЭБ, зафиксированное в определенный момент времени.

Файл исходных данных описанной выше структуры можно сформировать на основе анализа запросов к ЭБ (анализа соответствующих файлов аудита).

Для более детальной формализации постановки задачи и подходов к ее решению введем следующие основные понятия и обозначения, используемые в работе [9]:

- Алфавит Σ – множество атомарных действий.
- След – последовательность атомарных действий за время сессии работы пользователя (последовательность элементов из алфавита Σ).
- Набор следов – набор из нескольких сессий работы пользователя (множество, элементами которого являются следы).
- Классификатор – функция $f : \Sigma^* \rightarrow B$, где Σ* – набор, состоящий из конечных следов, множество B = {0, 1}.
- Нормальная активность – действия пользователя, заведомо удовлетворяющие требованиям безопасности, т.е. те действия, которые не трактуются системой как вторжение.
- Аномальная активность – действия пользователя, направленные на осуществление полного копирования документов ЭБ.
- Тренировочный набор – набор следов нормальной активности.

3 Метод, основанный на построении Марковской цепи

Опишем метод построения Марковской цепи (МЦ), которая будет являться основой шаблона «нормального» поведения пользователя и соответствующего классификатора [9].

Пусть имеется алфавит ∑, множество всех конечных следов T* и тренировочный набор, составленный из нормальных следов $T_{tr} \in T^*$. По набору T_{tr} создается МЦ (шаблон «нормального» поведения), на основе которой строится классификатор следов, предназначенный для формирования решения о нормальности/аномальности действий пользователя.

Расширим алфавит ∑ специальной буквой, обозначающей пустой символ (действие): Ø. При построении МЦ зададим параметр – «окно» размера w . Отметим, что состояние в МЦ связано со следом длины w через алфавит ∑UØ, т. е. каждое состояние – набор из w символов алфавита ∑UØ. Переход – это пара состояний (s, s') определяющая переход из s в s'. Каждое состояние и переход также связаны со счетчиком количества переходов.

Операция shift(σ, x) сдвигает след σ влево и добавляет символ x в конец следа, т. е. shift(«aba», c) = «bac».

Начальное состояние МЦ определяется как след длины w, состоящий из нулевых символов, т. е. если w=3, то начальное состояние будет следом [Ø, Ø, Ø].

Операция next(σ) возвращает первый символ следа σ и сдвигает σ на одну позицию влево, т.е. next(«abcd») возвращает a и обновляет след до «bcd».

Для каждого следа $\sigma \in T_{tr}$, пока не обработаны все символы, входящие в алфавит, выполняются следующие шаги:

1. полагаем c = next(σ).
2. устанавливаем <следующее состояние> = shift(<текущее состояние>,c).
3. увеличиваем счетчики для состояния <текущее состояние> и перехода (<текущее состояние>, <следующее состояние>).
4. обновляем <текущее состояние> до значения <следующее состояние>.

После того как все следы из набора T_{tr} обработаны, каждое состояние и переход имеют связанные с ними целые положительные числа – счетчики. Вероятность перехода из состояния s в состояние s' (P(s,s')) полагается равной N(s,s')/N(s), где N(s,s') и N(s) счетчики, связанные с переходом (s,s') и s соответственно. Другими словами, вероятность перехода (s,s') – отношение частоты переходов (s,s') к частоте исходящего состояния s в наборе T_{tr}.

По построению P является корректной мерой, т.е. выполняется следующее соотношение для всех состояний s:

$$\sum_{s \in SUCC(s)} P(s,s') = 1$$

Здесь succ(s)={s′: в построенной МЦ существует переход (s,s’)} определяет набор преемников s.

Рассмотрим пример, представленный в работе [9].

Пусть тренировочный набор T_{tr} равен {aabc, abcbc}.

Структура МЦ, полученная в результате сканирования всех следов тренировочного набора Ttr показана на Рис. 1. На этом рисунке также указаны счетчики, связанные с состояниями и переходами.

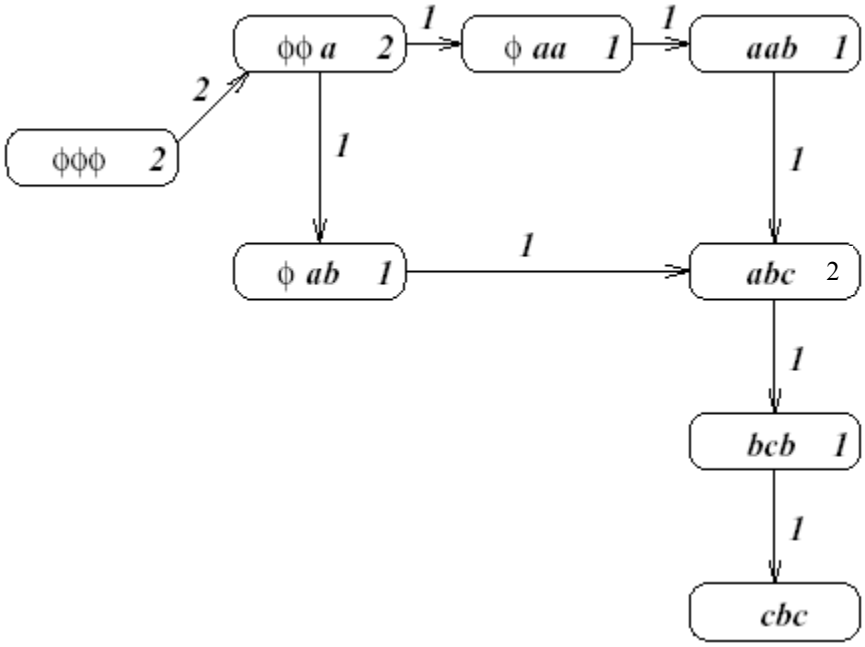

Рис. 1. Структура Марковской цепи

Построенная подобным образом МЦ представляет собой шаблон «нормального» поведения, который создается для каждого зарегистрированного в системе пользователя. Для фиксации отклонений действий пользователей от шаблонного поведения необходимо также построить классификатор поведения, задачей которого является проверка каждого нового зафиксированного действия пользователя на нормальность/аномальным.

4 Построение классификатора поведения

Обозначим за s_0 след начального состояния, т.е. след, состоящий из пустых символов алфавита Ø. След длины w, ассоциированный с состоянием s обозначается как σ(s).

Рассмотрим след α є Σ*, α[i] обозначает i-й символ следа α. Пусть начальный след β_0 будет равен σ(s_0), т.е. следу начального состояния s_0, состоящему из w пустых символов. След после сканирования первого символа α[1] будет β_1 = shift(β_0, α[1]). След β_k получается после сканирования k-го символа и рекурсивно определяется как shift(β_{k-1}, α[k]). Следовательно, след α определяет последовательность следов β_0, ..., β_m (где каждый след β_i длины w и m=|α|).

Определим метрику μ(α), соответствующую следу α. Эта метрика будет базироваться на построенной ранее МЦ и будет вычисляться итеративно. Изначально положим X и Y равными 0.0 и i = 0. Пока i ≠ m мы выполняем следующие шаги:

- Для следов β_i и β_{i+1} рассматриваются два варианта:

(Вариант А): $\beta_i \rightarrow \beta_{i+1}$ существующий переход в МЦ.

Если два состояния s и s' в МЦ такие, что σ(s) = β_i и σ(s') = β_{i+1}, тогда обновляем X и Y через функции-параметры F и G согласно следующим правилам:

Y = Y + F(s, (s,s'));

X = X + G(s, (s,s'));

(Вариант Б): β_i или β_{i+1} не являются существующими состояниями МЦ.

Если $\beta_i \rightarrow \beta_{i+1}$ невозможный переход в МЦ, тогда обновляем X и Y по следующим правилам (параметр – число Z):

Y = Y + Z;

X = X + 1;

- Увеличиваем i до i+1.

Метрика μ(α) определяется как Y/X в конце процедуры, описанной выше. Только что описанная процедура определяет функцию μ: Σ* → ℜ. Интуитивно понятно, что метрика μ(α) оценивает насколько хорошо МЦ предсказывает след α, т.е. малое значение μ(α) говорит о том, что МЦ предсказывает след α хорошо. Отметим, что μ параметризована функциями F, G и числом Z. Различный выбор F и G будет изменять значение классификатора.

Пусть дан порог r є ℜ. Классификатор *f* может быть построен из метрики μ следующим образом:

$$f(a) = \begin{cases} 1, \mu(a) > r \\ 0, иначе \end{cases}$$

Т.е. след α классифицируется как аномальный, если метрика μ на этом следе превышает пороговое значение r.

Нетрудно проверить, что для примера, представленного на Рис.1, классификатор, построенный из метрики с параметрами F(s,

(s,s'))≡1, G(s, (s,s'))≡1, Z=2 с порогом r=1, укажет на след {aacb} как на аномальный.

Однако описанный выше метод построения классификатора и шаблона «нормального» поведения пользователя имеет уязвимость, позволяющую потенциально провести атаку на защищаемую систему [4].

Суть этой уязвимости покажем на примере. Рассмотрим следующую схему проведения атаки. Пусть атака состоит из некоторых действий $\{a_1a_2\ldots a_k\}$ и существует последовательность действий $\{b_1b_2\ldots b_M\}$, где $M \geq W$, являющаяся законченным блоком действий, не несущих опасности (например, установление соединения с удаленным сервером, чтение новостей и др.). Если на этапе сбора тренировочных данных в действиях пользователя были следы вида $\{b_1b_2\ldots b_M\mathbf{a_1}b_1b_2\ldots b_M\}$, $\{b_1b_2\ldots b_M\mathbf{a_2}b_1b_2\ldots b_M\}$, …, $\{b_1b_2\ldots b_M\mathbf{a_k}b_1b_2\ldots b_M\}$, то последовательность действий $\{b_1b_2\ldots b_M\mathbf{a_1}b_1b_2\ldots b_M\mathbf{a_2}\ldots b_1b_2\ldots b_M\mathbf{a_k}\}$, содержащая все действия атаки, будет оценена описанным выше классификатором как безопасная, т.к. каждое из подмножеств конечного следа мощностью M будет встречаться в построенной цепи МЦ. В то же время, т.к. атакующие действия $\{a_1a_2\ldots a_k\}$ перемежаются с не влияющими на систему блоками законченных действий, атака достигнет своей цели. То есть, при некоторых условиях злоумышленник получает возможность «спрятать» проводимую на компьютерную систему атаку за следами нормальной активности.

Для устранения описанной выше уязвимости и повышения эффективности моделирования «нормального» поведения пользователя предлагается использовать принцип «безопасного блока» [3].

Суть «принципа безопасного блока» заключается в том, что последовательности действий, встречающиеся в следах пользователя с частотой, превышающей некоторый порог, являются безопасными вне зависимости от того, какие действия (или последовательности действий) предшествовали или следовали за данной последовательностью и не рассматриваются при построении профиля пользователя. «Безопасные блоки» выделяются из исходных данных поведения пользователя с помощью алгоритмов Data Mining. Таким образом, при построении МЦ из шаблона «нормального» поведения на этапе конструирования будут исключены «безопасные блоки», не несущие угрозы безопасности системы в целом.

С учетом описанного выше принципа алгоритм построения классификатора поведения пользователя будет выглядеть следующим образом:

1. используя алгоритмы Data Mining (например, алгоритм Apriori, представленный в работе [2]), выделяются и исключаются «безопасные блоки» – последовательности действий, встречающиеся в тренировочных данных о поведении пользователя с частотой, превышающей некоторый параметр α;
2. на основе обработанного множества тренировочных данных (не содержащих «безопасных блоков») строится Марковская цепь (по описанному ранее методу), являющаяся профилем «нормального» поведения.

Алгоритм классификации действия пользователя x_i дополняется следующим модулем обработки «безопасных блоков»:

- если действие x_i завершает «безопасный блок» – возвращается значение N(x) за текущее действие и за все отложенные штрафы;
- если последовательность $x_{i-k}\rightarrow x_{i-k+1}\rightarrow\ldots\rightarrow x_i$ принадлежит какому-либо «безопасному блоку» – возвращается значение $D(x_i)$;
- если переход $x_{i-1} \rightarrow x_i$ принадлежит МЦ – возвращается значение $N(x_i)$ за текущее действие и за все отложенные штрафы;
- иначе – возвращается значение A(x) за текущее действие и за все отложенные штрафы.

Возвращаемые модулем обработки «безопасных блоков» значения представляют собой двухкомпонентные вектора, которые складываются с вектором (X, Y) и имеют следующий смысл: $A(x)=(a_1(x),a_2(x))$ – аномальное действие (вариант Б исходного алгоритма), $N(x)=(n_1(x),n_2(x))$ – нормальное действие (вариант А исходного алгоритма) и $D(x)=(d_1(x),d_2(x))$ – отложенный штраф.

Отложенный штраф означает, что на текущем шаге принятие решения о нормальности/аномальности действия откладывается и будет принято на основе анализ дальнейшего поведения пользователя.

След $\sigma=\{x_1x_2\ldots x_n\}$ считается аномальным, если метрика μ на каком-то шаге превышает значение r – порог обнаружения.

Как известно, в процессе эксплуатации системы обнаружения вторжений, основанные на аномальном подходе, могут совершать следующие ошибки при классификации поведения пользователей:

- обнаружение аномального поведения (согласно используемому профилю пользователя), которое не является атакой, и отнесение его к классу атак (ошибка типа false positive);
- пропуск атаки, которая не попадает под определение аномального поведения (ошибка типа false negative).

Также важной характеристикой работы системы являются затраты времени на обнаружение вторжения, которые показывают как много аномальных действий сможет совершить

злоумышленник до того, как будет обнаружен системой обнаружения вторжений.

Оптимальное соотношение параметров построения МЦ, «безопасных блоков» и классификатора обнаружения вторжения для уменьшения количества ошибок типов false negative и false positive, а также для увеличения скорости обнаружения вторжения, определяется при проведении испытаний системы на тестовых наборах следов.

Система защиты от несанкционированного полного копирования электронного фонда ЭБ может быть встроена в систему обеспечения доступа к ресурсам следующим образом. В процессе работы пользователя в системе каждое его действие (поиск, загрузка, копирование) с цифровым документом вызывает обращение к подсистеме классификации поведения пользователя. Если очередное действие пользователя признается аномальным и метрика μ превысила на текущем следе некоторый заданный порог r, то автоматически генерируется сигнал об обнаружении вторжения и пользователю блокируется доступ к ресурсам ЭБ.

Описанный выше подход к защите ЭБ от полного несанкционированного копирования можно назвать бихевиористским (т. е. поведенческим), т.к. основным критерием зависимости (или связи) документов в ЭБ – которая и определяет нормальность/аномальность двух последовательных обращений к различным цифровым документам – считается поведение пользователей, выражающееся в осуществленных и неосуществленных переходах между документами.

5 Заключение

В работе рассматривается задача построения системы защиты ЭБ от несанкционированного полного копирования документов на основе применения аномального (статистического) подхода. Для апробации предлагаемых решений планируется разработка и тестирование исследовательского прототипа системы обнаружения вторжений (защиты от несанкционированного полного копирования документов) для электронной библиотеки научных информационных ресурсов КарНЦ РАН (http://dl.krc.karelia.ru) [1]. С этой целью предполагается решение следующих основных задач:

- разработка модуля сбора информации, связанной с работы системы обнаружения вторжений, а также структур данных для оптимального представления и хранения полученной информации;
- создание наборов следов «нормальной» активности пользователей и построение на их основе шаблонов «нормального» поведения пользователей;
- разработка системы тестов для оценки характеристик системы, проведение серии вычислительных экспериментов для подбора оптимальных значений параметров построения системы с целью уменьшения количества возможных ошибок типа false positive и false negative и снижения среднего времени обнаружения атаки.

Также представляется целесообразным провести исследования, связанные с возможностью построения шаблонов «нормального» поведения пользователей на основе разрабатываемой для ЭБ КарНЦ РАН онтологии.

Литература

[1] Вдовицын В.Т., Лебедев В.А., Луговая Н.Б., Сорокин А.Д., Старкова В.Г. Разработка и развитие технологии публикации и поиска документов в электронных коллекциях. //Труды Восьмой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”.Суздаль, Россия 17-19 октября 2006 г. – Ярославль: Ярославский государственный университет им. П.Г. Демидова, 2006. С. 162-167.

[2] Венкатеш Ганти, Йоханнес Герке, Раджу Рамакришнан. Добыча данных в сверхбольших базах данных. http://www.osp.ru/os/1999/09-10/053.htm

[3] Ивашко Е. Е. Применение методов Data Mining и Марковских цепей для построения классификатора системы обнаружения вторжений. //Сборник тезисов международной научно-практической Интернет - конференции “Современные проблемы и пути их решения в науке, транспорте, производстве и образовании”. Научно-исследовательский проектно-конструкторский институт морского флота Украины («УКРНИИМФ») и Одесский национальный морской университет. 2005 г.

[4] Ивашко Е. Е. Уязвимость классификатора системы обнаружения вторжений, основанного на Марковских цепях //Сборник тезисов международной научно-практической Интернет - конференции “Современные проблемы и пути их решения в науке, транспорте, производстве и образовании”. Научно-исследовательский проектно-конструкторский институт морского флота Украины («УКРНИИМФ») и Одесский национальный морской университет. 2005 г.

[5] Информационный портал "Российские электронные библиотеки". http://www.elbib.ru/

[6] А. А. Кузнецов. Обеспечение защищенного обмена информацией при организации работы с фондом служебного пользования электронной библиотеки. //Труды Шестой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL2004. Пущино, Россия 29 сентября – 1 октября 2004 г. С.175-184.

[7] Е. А. Негуляев, Е. А. Охезина. Создание и сбор полнотекстовых электронных ресурсов в университетской библиотеке // Электронные библиотеки, 2003, Том 6, Выпуск 5. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part5/NO

[8] Условия пользования научной библиотекой РФФИ. http://www.rsci.ru/MoreInfo.html?MessageID=498

[9] S. Jha, K. Tan, R.A. Maxion. Markov Chains, Classifiers and Intrusion Detection //Computer Security Foundations Workshop (CSFW), June 2001.

[10] Alexandros Koulouris, Sarantos Kapidakis. Considerations on policies of university digital collections //Труды Шестой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL2004. Пущино, Россия 29 сентября – 1 октября 2004 г. С. 159-168.

[11] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok. A Data Mining Framework for Intrusion Detection System //Computer Science Department of Columbia State University. June 30, 1998.

[12] Yun Xu, Renjin He. Database construction and copyright protection // Global Digital library development in the new millennium. Fertile ground for distributed cross-disciplinary collaboration. Tsinghua university press, Beijing, China, 2001. P. 545-546.

The defensive system against unauthorized documents-copying of the digital libraries development

E. Ivashko

Karelian Research Center RAS
Institute of Applied Mathematical Research
Petrozavodsk
ivashko@krc.karelia.ru

Annotation

In this article we consider problems of applying statistical anomaly detection algorithm to preserve digital libraries from unauthorized full-scale documents copying. We propose modernized intrusion detection technique based on Markov chains for creating classifiers and patterns of normal behavior.

* Работа поддержана грантом РФФИ № 05-07 -90077