Система обнаружения полного несанкционированного копирования документов электронных библиотек♣

© Е.Е. Ивашко, Н.Н. Никитина

Учреждение Российской академии наук
Институт прикладных математических исследований КарНЦ РАН,
г. Петрозаводск
{ivashko, nikitina}@krc.karelia.ru

Аннотация

В работе представлены новые результаты, полученные в ходе исследований по разработке системы обнаружения полного несанкционированного копирования документов электронных библиотек. Система основана на моделировании поведения пользователей с помощью Марковских цепей. Приведены основные результаты серии экспериментов, направленных на выбор наилучших значений параметров и классификатора модели.

1. Введение

Полнотекстовые электронные библиотеки составляют важную часть российской и мировой научной инфраструктуры. Результаты передовых научных исследований представлены в статьях, собранных в таких общеизвестных электронных библиотеках как, например, eLibrary.RU [1], SpringerLink [2], ScienceDirect [3] и других.

При эксплуатации электронных библиотек перед владельцами возникают проблемы сохранения уникальности электронного ресурса и гарантии соблюдения авторских прав. При этом, в силу трудоемкости создания качественного контента и низкой стоимости воспроизводства цифровых копий, указанные проблемы являются наиболее острыми. При получении неограниченного доступа к содержимому электронной библиотеки, злоумышленник может, например, скопировать все электронные документы и воспроизвести ресурс под другим доменным именем. Новый ресурс может быть использован для получения прибыли в обход интересов правообладателей (см., например, в [4]), совершения мошеннических действий и т. п. Таким образом, проблемы сохранения уникальности электронного ресурса и гарантирования соблюдения авторских прав непосредственно связаны с задачей защиты от полного несанкционированного копирования цифровых документов электронной библиотеки. При этом под полным несанкционированным копированием здесь и далее понимается получение электронных копий всех или большой части цифровых документов ЭБ без разрешения ее владельцев.

Законодательства развитых стран мира в той или иной мере защищают базы данных цифровых документов электронных библиотек от полного несанкционированного копирования. Согласно законодательству России, база данных охраняется авторским правом, если она является результатом творческой деятельности по подбору и/или расположению включенных в нее материалов. Одновременно с этим база данных может охраняться смежным правом, которое признается за ее изготовителем независимо от наличия авторских прав на эту базу данных. При этом исключительное право распространяется на те базы данных, создание которых требует существенных финансовых, материальных, организационных или иных затрат (см. [5-8]). В Европейском союзе существует специальный правовой режим *sui generis*, в соответствии с которым производитель базы данных вправе запрещать извлечение и/или повторное использование совокупности или существенной части содержания базы данных (см. [5,8-10]). В США ограничения на копирование значимой части базы данных зачастую прописываются в контракте и защищаются в рамках контрактного права [5, 8, 11].

В этой связи можно отметить факты судебного преследования организаций и частных лиц, которые в нарушение пользовательского договора скачивают большое количество документов электронной библиотеки (см., например, [22,23]).

Однако помимо юридической защиты, большое значение имеет использование технических средств защиты от полного несанкционированного копирования. Один из наиболее популярных методов — это ограничение числа документов, загруженных пользователем в определенный период

Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

времени. Такие ограничения, как правило, прописываются также в лицензионных договорах и правилах доступа к электронным библиотекам (см., например, [12] и [13]). Другой способ — защита от автоматического скачивания электронных документов с помощью различных реализаций CAPTCHA [14].

Тем не менее, указанные методы защиты от полного несанкционированного копирования при относительной простоте реализации имеют существенные недостатки. Например, при ограничении интенсивности скачивания документов точное время, необходимое для полного копирования документов определяется как

$$T = \frac{N_{docs}}{Accs \times F_{docs}}$$

где T — это время, которое необходимо затратить на скачивание всей коллекции документов; N_{docs} — общее число документов коллекции; $Accs$ — число учетных записей, доступных злоумышленнику и F_{docs} — число документов, которое разрешается скачать в единицу времени.

Методы, основанные на использовании CAPTCHA, также не гарантируют защиты, так как параллельно с развитием средств автоматического различения программ, не менее успешно развиваются и технологии их автоматизированного преодоления.

Разработка и внедрение эффективных механизмов защиты от полного несанкционированного копирования позволит сохранить уникальность общедоступных ЭБ и гарантировать защиту авторских прав. Это даст возможность одним ЭБ открыть свой доступ более широкому кругу читателей, а другим — более эффективно следить за соблюдением договоров с авторами и издательствами.

Целью представленной работы является разработка интеллектуальной системы обнаружения полного несанкционированного копирования документов, основанной на аномальном статистическом подходе.

В представленной статье описаны новые результаты, полученные в ходе проведения серии экспериментов, направленных на выбор наилучших значений параметров и классификатора модели.

2. Результаты предыдущих этапов исследований

При разработке системы обнаружения полного несанкционированного копирования мы полагаемся на следующие гипотезы:

- все цифровые документы электронной библиотеки, скачиваемые обычными пользователями, семантически связаны между собой;
- документы, скачиваемые злоумышленником при полном несанкционированном копировании, имеют слабую семантическую связь.

Таким образом, имея возможность определять, имеют ли между собой семантическую связь документы из определенного набора, можно выявлять и попытки несанкционированного полного копирования.

Для решения этой задачи мы используем «поведенческий» подход — семантические связи между документами определяются на основе анализа поведения обычных пользователей (для этого создается шаблон «нормального» поведения). Моделирование «нормального» поведения производится с помощью однородной Марковской цепи. Аппарат Марковских цепей широко применяется при моделировании поведения, построении «персонализированных» систем и систем обнаружения вторжений (см., например, [18,20,21]).

Значимое отклонение действий пользователя от «нормального» поведения считается проявлением полного несанкционированного копирования.

В ходе предыдущих этапов исследований были получены следующие основные результаты:

- сформулирована постановка и разработана базовая математическая модель задачи обнаружения полного несанкционированного копирования документов электронных библиотек (полученные результаты были представлены на конференции RCDL-2007 [15]);
- разработан прототип системы обнаружения полного несанкционированного копирования; на основе проведенных экспериментов обоснована применимость предлагаемого подхода к поставленной задаче (RCDL-2009 [16]);
- проведены первые эксперименты по оценке эффективности классификатора поведения пользователей и зависимостей между параметрами модели (RCDL-2010 [17]).

Ниже представлена более подробная информация о результатах исследований, проведенных на предыдущих этапах.

2.1 Моделирование поведения пользователя

Основой шаблона «нормального» поведения пользователя является однородная Марковская цепь (МЦ), построенная по записям поведения пользователей ЭБ.

Марковская цепь — это случайный процесс $X(t)$, определенный на отрезке времени $[0,T]$ с дискретным пространством состояний $S=\{1, 2, ..., k\}$, обладающий свойством

$$P(X_n=i_n \mid X_0=i_0, ..., X_{n-1}=i_{n-1}) = P(X_n=i_n \mid X_{n-1}=i_{n-1})$$

(или «будущее при заданном настоящем не зависит от прошлого»).

Опишем кратко метод построения Марковской цепи на основе данных об обращении пользователей к документам электронной библиотеки [15,18].

Будем называть *следом* зафиксированную последовательность обращений пользователя к

документам электронной библиотеки (т. е. лог сессии работы пользователя).

Алфавит ∑ атомарных действий составляет список доступных документов ЭБ; *T** — это множество всех конечных следов и $T_{tr} \in T^*$ — тренировочный набор, составленный из заведомо нормальных следов.

Расширим алфавит ∑ специальным символом *Ш*. При построении МЦ задается параметр – «*окно*» размера *w*. Состояние в МЦ связано со следом длины *w* через алфавит ∑*UШ*, т. е. каждое состояние кодируется набором из w символов алфавита ∑*UШ*. Переход – это пара *(s, s′)*, определяющая в МЦ переход из состояния *s* в *s′*. Состояния и переходы связаны со счетчиками количества переходов.

Операция *shift(y, x)* сдвигает след *y* 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 *[Ш, Ш, Ш]*.

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

Алгоритм построения МЦ следующий (согласно [18]).

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

- полагаем *c = next(y)*.
- устанавливаем *<следующее состояние> = shift(<текущее состояние>,c)*.
- увеличиваем счетчики для состояния *<текущее состояние>* и перехода *(<текущее состояние>, <следующее состояние>)*.
- 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.

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.

Построенная по такому алгоритму МЦ представляет собой шаблон «нормального» поведения.

На рис. 1 показан пример МЦ, построенной по набору T_{tr}*={aabc, abcbc}*.

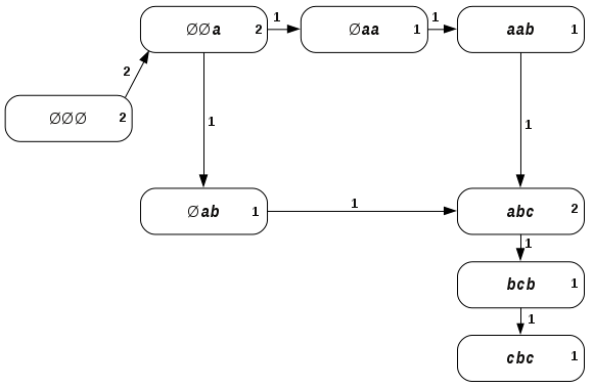

Рис. 1. Структура Марковской цепи

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

Рассмотрим след *б* $\in$ *У**; *б[i]* обозначает *i*-й символ следа *б*. Пусть начальный след $в_0$ будет равен $y(s_0)$, т.е. следу начального состояния s_0, состоящему из *w* пустых символов. След после сканирования первого символа *б[1]* будет $в_1$ = *shift*($в_0$, *б[1]*). След $в_k$ получается после сканирования *k*-го символа и рекурсивно определяется как *shift*($в_{k-1}$, *б[k]*). Следовательно, след *б* определяет последовательность следов $в_0, ..., в_m$ (где каждый след $в_i$ длины *w* и $m=|б|$).

Определим метрику *м(б)*, соответствующую следу *б*. Эта метрика будет основана на построенной ранее МЦ и будет вычисляться итеративно. Изначально положим *X* и *Y* равными *0.0* и *i = 0*. Пока $i \neq m$, мы выполняем следующие шаги:

- Для следов $в_i$ и $в_{i+1}$ рассматриваются два варианта:

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

Если два состояния *s* и *s′* в МЦ такие, что $y(s) = в_i$ и $y(s') = в_{i+1}$, 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): $в_i$ или $в_{i+1}$ не являются существующими состояниями МЦ.

Если $в_i \rightarrow в_{i+1}$ невозможный переход в МЦ, тогда *X* и *Y* обновляются согласно следующим правилам:

$$Y = H(Y);$$
$$X = L(X);$$

- Увеличиваем *i* до *i+1*.

Метрика *м(б): У* → Я* определяется как *Y/X* в конце представленной процедуры. Интуитивно понятно, что метрика *м(б)* оценивает насколько хорошо МЦ предсказывает след *б*, т.е. малое значение *м(б)* говорит о том, что МЦ предсказывает след *б* хорошо. Отметим, что *м* параметризована функциями *F*, *G* и числом *Z*. Различный выбор *F* и *G* будет изменять значение классификатора.

Пусть дан порог $r \in Я$. Классификатор *ŕ* может быть построен на основе метрики м следующим образом:

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

Т.е. след *б* классифицируется как аномальный, если метрика *м* на этом следе превышает пороговое значение *r*.

Нетрудно проверить, что для примера, представленного на рис. 1, классификатор, построенный из метрики *м(б)* 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.

В работе [16] представлены результаты первых экспериментов, проведенных в рамках разработки системы обнаружения полного несанкционированного копирования. Несмотря на некоторые ограничения, (связанные, в частности, с отсутствием в лог-файле информации о пользователях), полученные результаты позволили сделать вывод о применимости аномального подхода в обнаружении вторжений к обнаружению полного несанкционированного копирования документов ЭБ:

- на основе анализа поведения пользователей возможно автоматически выявлять семантические связи между электронными документами;
- возможно автоматическое выявление последовательностей обращений, противоречащих семантическим связям между документами.

2.2 Исследование влияния параметров на характеристики модели

Одним из наиболее важных параметров при построении классификатора поведения является значение порога, при превышении которого последовательность действий классифицируется как аномальная.

Цель работы [17] заключалась в исследовании влияния параметров модели (значения порога) на соотношение ошибок типа *false negative* (классификация аномального поведения как нормального) и *false positive* (классификация нормального поведения как аномального), а также на среднее время до обнаружения аномального поведения.

На рис. 2 представлены графики числа ошибок типа false negative и доли аномальных сессий в зависимости от значения порога (порог 1.8 соответствует примерно 5 аномальным действиям). Из рисунка видно, что при снижении числа ошибок типа false negative повышается число сессий, классифицированных как аномальные.

На рис. 3 показана зависимость среднего времени до обнаружения аномалии от значения порога. Естественно, с увеличением порога возрастает и среднее время до обнаружения аномалии.

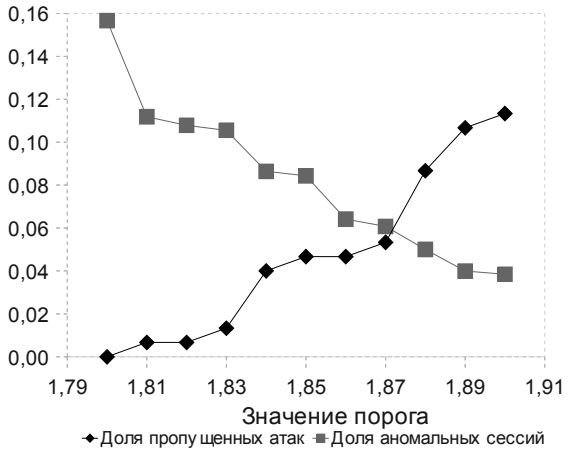

Рис. 2. Доля пропущенных атак и аномальных сессий в зависимости от значения порога

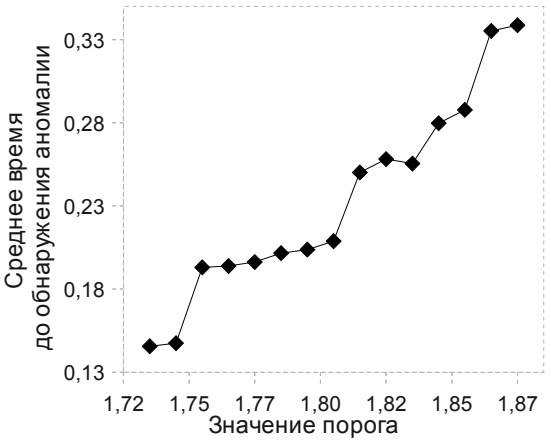

Рис. 3. Среднее время до обнаружения аномалии в зависимости от значения порога

На следующем этапе работы было запланировано проведение новой серии экспериментов для выяснения ключевых характеристик классификатора поведения пользователей.

3. Эксперименты

В данном разделе описаны новые эксперименты, проведенные в ходе разработки системы обнаружения полного несанкционированного копирования. При реализации экспериментов использовалась программная система, созданная на предыдущих этапах работы (см. [15-17]).

3.1 Исходные данные

Исходными данными для проведения экспериментов послужили лог-файлы доступа к документам Электронной библиотеки Республики Карелия [19], собранные за период с сентября 2004 г. по январь 2011 г.

Для проведения экспериментов были отобраны сессии работы зарегистрированных пользователей,

содержащие от 5 до 50 обращений к документам электронной библиотеки. Сессией работы считалась последовательность обращений к цифровым документам, в которой время, прошедшее между двумя обращениями, не превышало 12 часов. Всего было получено 10393 такие сессии, принадлежащие 5561 пользователю и содержащие обращения к 2071 уникальному документу электронной библиотеки. Общее количество обращений к документам в выделенных сессиях составило 109847.

3.2 Результаты экспериментов

Для проведения экспериментов был создан набор заведомо аномальных псевдосессий, содержащих переходы между документами из различных тематических разделов библиотеки. Эти данные, будучи заведомо аномальными, использовались для оценки эффективности системы обнаружения несанкционированного полного копирования документов. Все данные, полученные из Электронной библиотеки Республики Карелия, считались заведомо нормальными и использовались, во-первых, для создания шаблона «нормального» поведения пользователя, во-вторых, для оценки числа ложных срабатываний.

Первая серия экспериментов была направлена на исследование свойств линейного классификатора, используемого в работе [18] и представленного в общем виде в разделе 2 данной работы.

На рис. 4 представлены графики, демонстрирующие соотношение числа обнаруженных атак и числа ложных срабатываний при изменении значения порога от 0,0 до 4,0 при различных значениях Z (здесь Z используется как «штраф» за осуществление аномального перехода: L(X)=X+Z, см. раздел 2).

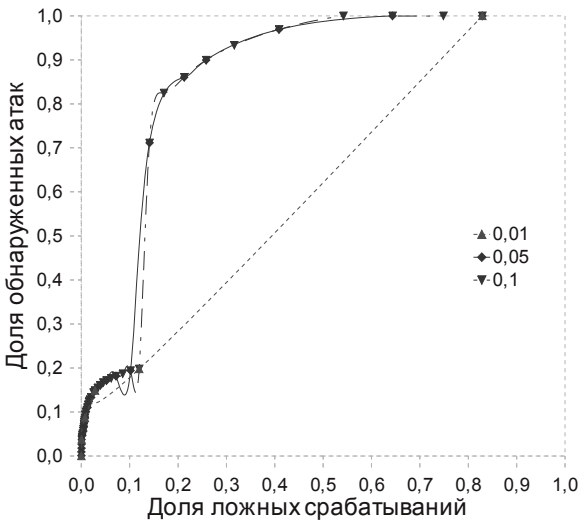

Рис. 4. Соотношение числа обнаруженных атак и ложных срабатываний для линейного классификатора

Как показали эксперименты, наилучшие результаты достигаются при Z=0,1 и пороге r=0,9 — в этом случае обнаруживается более 80% атак, но число ложных срабатываний достигает 15%, что недопустимо для систем, использующихся на практике.

В целом, при проведении целого ряда экспериментов по варьированию параметров модели, было выявлено чрезмерно большое число ложных срабатываний. Исходя из этого, можно сделать вывод о неприспособленности классификатора, описанного в работе [18], для реализации в системах обнаружения полного несанкционированного копирования.

В ходе дальнейших экспериментов был исследован ряд классификаторов, отличных от представленных в работе [18]. Среди них наилучшую производительность показал классификатор, имеющий логарифмическую зависимость от частоты переходов между состояниями МЦ. Результаты моделирования соотношения числа обнаруженных атак и ложных срабатываний представлены на рис. 5.

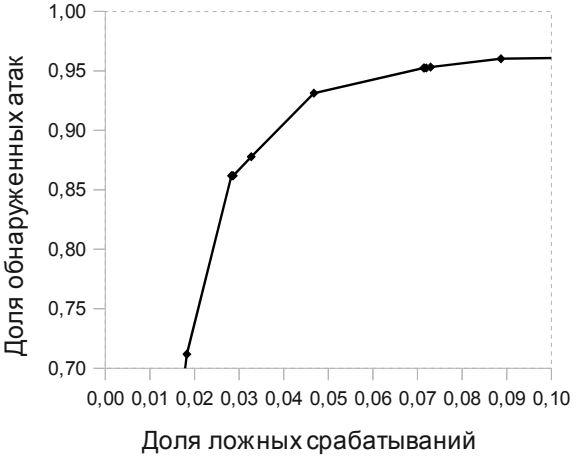

Рис. 5. Соотношение числа обнаруженных атак и ложных срабатываний для логарифмического классификатора

Таким образом, рассмотренный классификатор позволяет обнаружить порядка 95% атак при сравнительно небольшом числе ложных тревог.

3.3 Ресурсоемкость модели

Одним из наиболее важных практических вопросов при применении систем реального времени является вопрос ресурсоемкости.

На рис. 6 представлены графики временных затрат на построение шаблона «нормального» поведения в зависимости от размера «окна» модели и числа рассматриваемых сессий работы.

Как видно из графика, в худшем случае при 10393 сессиях (см. п. 3.1) даже при большом размере «окна» время построения модели «нормального» поведения на компьютере «офисной» конфигурации не превышало четырех минут. При этом затраты времени растут линейно при росте объемов исходных данных. Учитывая, что при работе системы такая модель создается лишь однажды, подобные временные затраты являются

приемлемыми на практике.

Другая важная характеристика — это объем оперативной памяти, необходимой для хранения шаблона «нормального» поведения. Согласно построению, число состояний МЦ может достигать значения

$$N^{w}(N+1),$$

где N — число уникальных документов ЭБ,
а w — размер «окна».

При этом для хранения каждого состояния требуется

$$24+4 \cdot w$$

байт памяти.

Таким образом, максимальный объем оперативной памяти в байтах, необходимый для хранения шаблона «нормального» поведения составляет

$$N^{w}(N+1) \cdot (24 + 4 \cdot w)$$

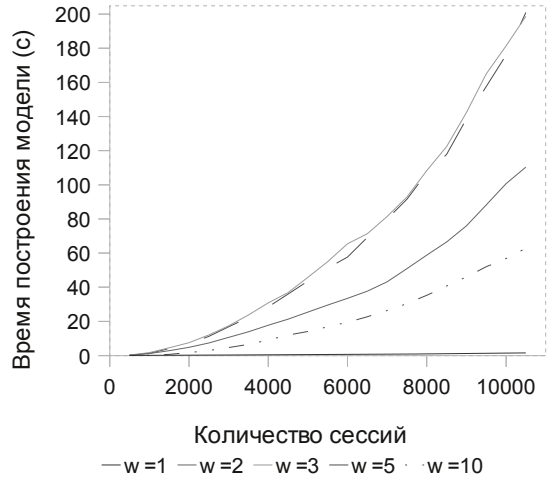

Рис. 6. Временные затраты на построение МЦ в зависимости от числа сессий и размера «окна»

Это значение быстро растет с увеличением размера «окна» и ростом числа уникальных документов, зафиксированных в лог-файлах.

Однако, как показали эксперименты, на практике число состояний составляет порядка 0,01 от максимального и этот коэффициент уменьшается с ростом числа документов. При построении шаблона «нормального» поведения по исходным данным, указанным в разделе 3.1, объем используемой для МЦ оперативной памяти составил 1,2 мегабайта, что является приемлемым для системы, интегрированной с электронной библиотекой.

Тем не менее, требуется провести оптимизацию алгоритма построения шаблона «нормального» поведения пользователя в целях снижения максимального объема необходимой оперативной памяти.

4. Заключение

В работе представлены результаты, полученные в ходе проведения экспериментов, направленных на исследование характеристик классификатора поведения пользователя и ресурсоемкости построения модели «нормального» поведения.

По результатам проведения экспериментов были сделаны следующие выводы:

1) классификатор, предложенный в работе [18] для систем обнаружения вторжений, не подходит для использования в практических системах обнаружения полного несанкционированного копирования. При этом класс логарифмических классификаторов может быть достаточно эффективным;

2) ресурсоемкость (время построения модели и объем занимаемой оперативной памяти) является приемлемой, однако требуется снижение порога максимально возможных затрат оперативной памяти.

Одним из недостатков статистического подхода к выявлению семантических связей между документами является необходимость накопления большого объема исходных данных. При этом вопрос о том, сколько именно записей сессий работы пользователей будет достаточно для построения эффективного шаблона «нормального» поведения, является сложным. Также возникают проблемы при обновлении шаблона, необходимом при добавлении новых документов в электронную коллекцию библиотеки. Решению этих задач может способствовать привлечение других методов выявления семантических связей между документами, например, использование онтологических моделей электронной коллекции.

В целом, учитывая сделанные ранее выводы о применимости подхода к поставленной задаче, необходимо продолжить работы по разработке системы обнаружения полного несанкционированного копирования, основанной на аномальном статистическом подходе.

Запланированы следующие перспективные направления работы:

1) дополнительные исследования логарифмического классификатора поведения пользователя, имеющего лучшие характеристики по соотношению числа обнаруженных атак и ложных срабатываний;

2) оптимизация структур данных, используемых для хранения шаблона «нормального» поведения пользователя с целью минимизации объема используемой оперативной памяти.

Кроме того, достигнуты предварительные договоренности о проведении совместной исследовательской работы по оценке разрабатываемого метода на исходных данных научной электронной библиотеки eLIBRARY.RU [1], а также об интеграции прототипа системы обнаружения полного несанкционированного копирования с программным обеспечением электронной библиотеки Республики Карелия [19].

Отметим также, что достоинством исследуемого

подхода является отсутствие привязки к типу содержимого электронной библиотеки — метод может быть модернизирован для защиты коллекций аудио- и видео-записей (например, для таких сайтов как RuTube.ru, YouTube.com и др.). При этом шаблоны «нормального» поведения могут быть либо едиными для всех типов контента, либо строиться и использоваться независимо.

Литература

[1] Научная электронная библиотека eLIBRARY.RU. http://www.elibrary.ru

[2] SpringerLink — полнотекстовая база данных книг, журналов и других изданий, выпускаемых издательством Springer. http://www.springerlink.com

[3] Электронная коллекция публикаций по результатам научных исследований ScienceDirect. http://www.sciencedirect.com

[4] «Дело о плагиате в российском интернете» www.medialaw.ru/publications/zip/62/ch2.htm

[5] Боровская Е.А., Ермакович С.Л., Кудашов В.И., Лосев С.С., Успенский А.А. Правовая охрана компьютерных программ и баз данных. Белорусский республиканский центр трансфера технологий. http://www.ictt.by

[6] Гражданский кодекс, часть 4. Глава 70 Авторское право. http://www.gk-rf.ru/glava70

[7] Гражданский кодекс, часть 4. Глава 71 Права, смежные с авторскими. http://www.gk-rf.ru/glava71

[8] Вайшнурс А. А. Современность и перспективы правовой охраны баз данных в России, США и Европейском союзе. http://www.ruvento.com/files/file/Evolution_of_database_protection_in_Russia_USA_and_EU.pdf

[9] Директива ЕС №96/9/ЕС от 11 марта 1996 г. о правовой охране баз данных. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML

[10] Ханина К. В. Воздействие развития информационных технологий на формирование системы авторского права ЕС. Доклад на V Международной конференции «Право и Интернет», 2003 г. http://www.ifap.ru/pi/05/sr02.doc

[11] M. Smith. A Comparison of the Legal Protection of Databases in the United States and EU: Impications for Scientific Research. Social Science Research Network, 2010. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1613451

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

[13] Лицензионное соглашение сайта проекта eLIBRARY.RU ООО «Научная электронная библиотека». http://elibrary.ru/agreement.asp

[14] CAPTCHA - http://ru.wikipedia.org/wiki/Captcha

[15] Ивашко Е. Е. Построение системы защиты электронных библиотек от полного несанкционированного копирования документов //Труды Девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Переславль, Россия, 15-18 октября 2007 г. — Переславль-Залесский: изд-во «Университет города Переславля», 2007. С. 300-306.

[16] Ивашко Е. Е., Никитина Н. Н. Опыт построения системы защиты электронных библиотек от несанкционированного копирования документов //Труды Одиннадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Петрозаводск, Россия, 17-21 сентября 2009 г. — Петрозаводск: КарНЦ РАН, 2009. С. 443-447.

[17] Ивашко Е. Е., Никитина Н. Н. Аномальный подход к обнаружению полного несанкционированного копирования документов электронной библиотеки //Труды Двенадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Казань, Россия, 13-17 октября 2010 г. — Казань: Казан. Ун-т, 2010. С. 111-115.

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

[19] Электронная библиотека Республики Карелия. http://www.elibrary.karelia.ru

[20] D. Nicholas, P. Huntington, H. R. Jamali, A. Watkinson// The information seeking behaviour of the users of digital scholarly journals. Information Processing and Management, 42. 2006. P 1345-1365.

[21] E. Frias-Martinez, G. Magoulas, S. Chen, R. Macredie. Automated user modeling for personalized digital libraries// International Journal of Information Management, 26. 2006. P. 234-248.

[22] Copyright Lawsuit against Georgia State University. http://ourgeorgiahistory.com/ogh/Copyright Lawsuit against Georgia State University.

[23] Internet Activist Charged in M.I.T. Data Theft http://bits.blogs.nytimes.com/2011/07/19/reddit-co-founder-charged-with-data-theft/

A System to Detect the Large-Scale Copying of Documents from Digital Libraries

© Evgeny E. Ivashko, Natalia N. Nikitina

In this work we present the new results obtained during the implementation of the research devoted to development of the system to prevent large-scale copying of documents from digital libraries. The system is based on user's behavior modeling applying with the Markov model. We present main results of experiments aimed at selection of the best values for the parameters and the classifier of the model.

♣ Работа поддержана грантом ЗАО “Лаборатория Касперского” в рамках “Программы поддержки инновационных проектов”, 2011 г.