Аномальный подход к обнаружению полного несанкционированного копирования документов электронной библиотеки[i]

© Е.Е. Ивашко, Н.Н. Никитина

Учреждение Российской академии наук

Институт прикладных математических исследований КарНЦ РАН, г. Петрозаводск

ivashko@ krc.karelia.ru, nikitina@ krc.karelia.ru

Аннотация

Представлены новые результаты, полученные в ходе разработки системы обнаружения полного несанкционированного копирования документов электронной библиотеки. В основу системы положен аномальный подход к обнаружению вторжений и моделирование шаблонов «нормального» поведения пользователя с помощью Марковских цепей.

1 Введение

Полнотекстовые электронные библиотеки (ЭБ) являются важной частью информационной среды, сформированной широким распространением интернета. Документы, представляемые в электронном виде, формируют репутацию учреждений и граждан (например, электронные коллекции научных работ отдельных ученых и/или научных учреждений), способствуют ликвидации «цифрового неравенства», служат повышению эффективности деятельности научного сообщества.

Однако для успешного функционирования и развития ЭБ важны рост посещаемости, гарантия соблюдения авторских прав и уникальности ресурса. При этом, в силу открытости электронных коллекций и низкой стоимости воспроизводства цифровых копий, именно уникальность ресурса и авторские права подвергаются большой опасности. При получении открытого доступа к ЭБ потенциальный злоумышленник может, например, полностью скопировать сайт ресурса и воспроизвести его под другим именем. В дальнейшем созданный сайт может быть использован для получения прибыли в обход интересов правообладателей, совершения мошеннических действий, распространения вредоносных программ и т. д.

Авторское право защищает сайт, на котором функционирует ЭБ, как сборник произведений литературы, науки и искусства [1]. Под защиту попадает оригинальная структура расположения и размещения произведений, придуманная создателями ЭБ. Однако помимо юридической защиты большое значение имеет использование технических средств защиты от полного несанкционированного копирования. При этом под полным несанкционированным копированием понимается получение электронных копий всех или большой части цифровых документов ЭБ без согласия ее владельцев.

Проблема защиты электронных библиотек от полного несанкционированного копирования является достаточно новой. В настоящее время большое внимание уделяется борьбе с плагиатом (например, при помощи специальных средств обнаружения заимствований [2] или расстановки специальных «водяных знаков» в цифровых документах [3]). Однако задача сохранения уникальности (т. е. защиты от полного несанкционированного копирования), как правило, решается прямолинейно: ограничением круга пользователей, имеющих доступ к электронным коллекциям (как, например, в Dartmouth College Digital Library и Harvard University Library [4]) и/или ограничением числа документов, загруженных пользователем в единицу времени (упоминания о таком ограничении есть, например, на ресурсах [5] и [6]). Разработка и внедрение эффективных механизмов защиты от полного несанкционированного копирования позволят сохранить уникальность общедоступных ЭБ. Это даст возможность одним ЭБ открыть свой доступ более широкому кругу читателей, а другим – более эффективно следить за соблюдением договоров с авторами и издательствами.

В представленной работе рассматривается применение аномального статистического подхода в обнаружении вторжений к задаче обнаружения несанкционированного полного копирования документов ЭБ. Результаты, полученные на ранних этапах работы, были представлены на конференциях RCDL-2007 и RCDL-2009 [7, 8].

2 Обнаружение полного несанкционированного копирования на основе аномального подхода

В этом разделе представлено краткое описание модели, используемой при построении системы обнаружения несанкционированного полного копирования. В основу модели положен аномальный под-

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

ход в обнаружении вторжений (см., например, [9]). Более подробное формальное описание используемой модели можно найти в работе [7].

2.1 Аномальный подход к обнаружению вторжений

Аномальный подход хорошо зарекомендовал себя в системах обнаружения вторжений [9]. Основная идея этого подхода базируется на двух следующих гипотезах (получивших подтверждение на практике):

- в обычных информационных системах количество злоумышленников составляет не более долей процента от общего числа пользователей;
- действия злоумышленника значимо отличаются от действий обычных пользователей.

Как следствие, на основании действий обычных пользователей можно построить шаблон «нормального» поведения; тогда значимое отклонение от этого профиля свидетельствует об аномальном поведении, т. е. обнаружении злоумышленника.

При разработке системы, реализующей аномальный подход в обнаружении вторжений, возникают следующие основные задачи:

- построение шаблона «нормального» поведения пользователя;
- разработка классификатора, позволяющего отличить «нормальную» последовательность действий от аномальной;
- определение граничных значений характеристик классификатора для снижения вероятности появления ошибок классификации;
- периодическое обновление шаблонов «нормального» поведения.

В итоге эффективность работы системы обнаружения будет зависеть, в основном, от метода моделирования шаблона «нормального» поведения, способа построения классификатора и значений используемых при этом параметров. В данной работе для создания шаблона «нормального» поведения используется Марковская цепь, построенная по записям поведения обычных пользователей (т. е. по их обращениям к документам ЭБ). При классификации поведения должно учитываться количество отклонений от «нормального» шаблона: только достаточно большое число несоответствий шаблону свидетельствует об аномальном поведении.

В процессе эксплуатации системы обнаружения вторжений, основанные на аномальном подходе, могут совершать следующие ошибки классификации:

- ложная классификация поведения как аномального (ошибка типа false positive);
- ложная классификация поведения пользователя как «нормального» (ошибка типа false negative).

Также важной характеристикой работы системы являются затраты времени на обнаружение вторжения, которые показывают насколько много аномальных действий сможет совершить злоумышленник до того, как будет выявлен системой обнаружения вторжений.

2.2 Обнаружение несанкционированного полного копирования

Соблюдая договоры с издательствами и авторами, многие ЭБ вводят ограничение на количество документов, скачиваемых в единицу времени (час, день). Однако проблема такого подхода заключается в чрезмерной простоте защиты: зная допустимое число скачиваемых документов, можно точно указать, какое время необходимо для полного копирования всех документов ЭБ. Целью нашей работы является разработка на основе аномального подхода интеллектуальной системы защиты от несанкционированного полного копирования и оценка ее эффективности.

Мы модифицируем вторую гипотезу аномального подхода следующим образом:

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

Действительно, обращаясь за информацией в ЭБ, пользователь хочет получить ответ на определенный вопрос и/или подобрать список документов по определенной тематике. Даже с учетом смежных тем и варьирования тематики поиска в зависимости от вновь получаемой информации все документы, к которым обращается пользователь, как правило, будут взаимосвязаны.

Таким образом, имея возможность определять, связаны ли между собой документы из определенного набора, можно выявлять и попытки несанкционированного полного копирования. Однако задача выявления семантических связей между документами также является нетривиальной. Для решения этой задачи мы используем «поведенческий» подход – семантические связи между документами определяются на основе анализа поведения обычных пользователей (для этого и создается шаблон «нормального» поведения).

Исходными данными при создании системы обнаружения несанкционированного полного копирования на основе аномального подхода служат записи об обращениях пользователей к отдельным документам электронной библиотеки. Основная цель создаваемой системы защиты заключается в том, чтобы своевременно обнаружить пользователя ЭБ, осуществляющего несанкционированное полное копирование документов библиотеки. При этом необходимо распознавать и такие ситуации, когда пользователь маскирует свои действия путем перемешивания действий по полному копированию с обычной работой с ресурсом.

Для построения системы обнаружения несанкционированного полного копирования необходимы следующие наборы исходных данных:

- тренировочный набор заведомо безопасных для ЭБ действий пользователя (на его основе строится шаблон «нормального» поведения);
- тестовый набор заведомо безопасных дей-

ствий (используется для оценки числа ошибок типа false positive);

- тестовый набор заведомо аномальных действий (используется для оценки числа ошибок типа false negative).

Наборы исходных данных могут быть получены из зафиксированной за определенный срок истории посещений ЭБ пользователями. С помощью этих наборов данных параметры построения шаблона «нормального» поведения и классификатора подбираются таким образом, чтобы минимизировать число ошибок классификации.

В целом алгоритм построения системы обнаружения несанкционированного полного копирования выглядит следующим образом:

1. с помощью тренировочного набора данных строится Марковская цепь, представляющая собой шаблон «нормального» поведения пользователя;
2. на основе шаблона «нормального» поведения строится классификатор поведения пользователя;
3. с помощью тестовых наборов оценивается соотношение числа ошибок классификации;
4. шаги 1 – 3 повторяются с различными значениями параметров (построения шаблона «нормального» поведения и классификатора) до достижения приемлемого числа и соотношения ошибок классификации.

Система обнаружения несанкционированного полного копирования документов ЭБ может быть встроена в систему обеспечения доступа к ресурсам следующим образом. В ходе работы пользователя в системе каждое его обращение к цифровому документу фиксируется и вызывает обращение к подсистеме классификации поведения. Если (с учетом последнего зафиксированного действия) поведение пользователя признается аномальным, то, в зависимости от политики безопасности ЭБ, либо администратору отсылается уведомление, либо пользователю блокируется доступ к ЭБ.

Описанный выше подход к обнаружению полного несанкционированного копирования можно назвать поведенческим, т. к. основным критерием зависимости (или связи) документов в ЭБ – определяющей нормальность/аномальность серии последовательных обращений к различным документам – считается ранее зафиксированное поведение пользователей, выражающееся в осуществленных и неосуществленных переходах между документами.

3 Эксперименты

В этом разделе будут описаны эксперименты, проведенные в рамках разработки системы обнаружения полного несанкционированного копирования. Для проведения экспериментов была разработана программная система, выполняющая предварительную обработку файла исходных данных, построение профиля «нормального» поведения и проверку сессий работы пользователей на аномальность. В работе [8] представлены результаты первых экспериментов, показавших применимость описанного ранее аномального подхода к задаче обнаружения полного несанкционированного копирования.

Исходными данными для проведения экспериментов послужили лог-файлы доступа к Электронной библиотеке Республики Карелия [10] за период с июня 2007 г. по февраль 2009 г. включительно. Всего в ЭБ содержится порядка 1000 документов, из них за рассматриваемый период были зафиксированы обращения к примерно 700 документам.

Сессией работы пользователя считалась последовательность всех обращений к документам с конкретного IP-адреса. Всего в лог-файле содержится порядка 10 тысяч сессий работы пользователей.

Согласно модели, для построения шаблона «нормального» поведения необходимы три набора данных: тренировочный (заведомо нормальные данные) и два тестовых (для подбора оптимальных параметров). Шаблон «нормального» поведения, был построен на основе обращений к ЭБ, зафиксированных в период с июня 2007 г. по май 2008 г. включительно. Остальная часть лог-файла служила в качестве тестового набора данных. Дополнительно был создан набор заведомо аномальных сессий работы с ЭБ.

На рис. 1 представлен пример (фрагмент) «нормального» профиля, показывающий семантические связи между документами, выявленные на основе поведения пользователей ЭБ.

Естественно, что наиболее сильно оказываются связаны отдельные электронные документы и оглавления разделов. Однако наряду с этим связанными в профиле являются, например, такие документы как «Основные правовые системы современности» и «Обзор истории русского права». Из названий этих документов понятна их семантическая близость, что подтверждает исходный тезис о возможности выявления семантических связей между документами на основе анализа поведения пользователей ЭБ.

В работе [8], опираясь на результаты проведенных экспериментов, были сделаны два принципиальных вывода:

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

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

На рис. 2 представлены графики числа ошибок типа false negative и доли аномалий 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 последовательностей действий, классифицированных как аномальные.

На рис. 3 показана зависимость среднего времени до обнаружения аномалии от значения порога. Как и следовало ожидать, с увеличением порога

увеличивается и среднее время до обнаружения аномалии.

Следует заметить, что используемый нами подход не привязан к типу содержимого электронной библиотеки и легко может быть модернизирован для защиты коллекций аудио- и видеозаписей (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 независимо.

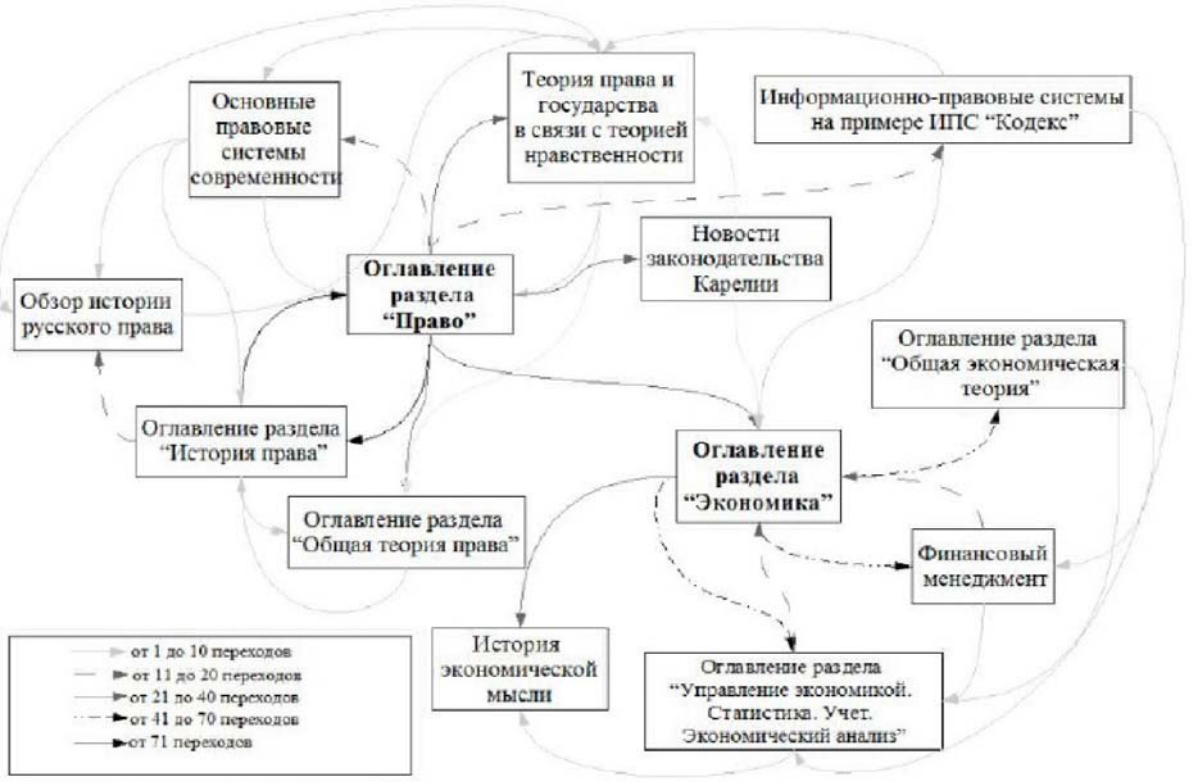

Рис. 1. Пример (фрагмент)

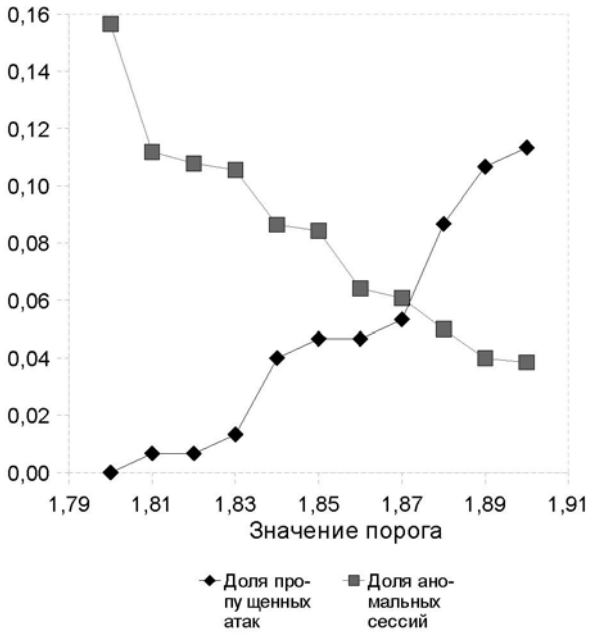

Рис. 2. Доля пропущенных атак и аномальных сессий в зависимости от значения порога профиля «нормального» поведения

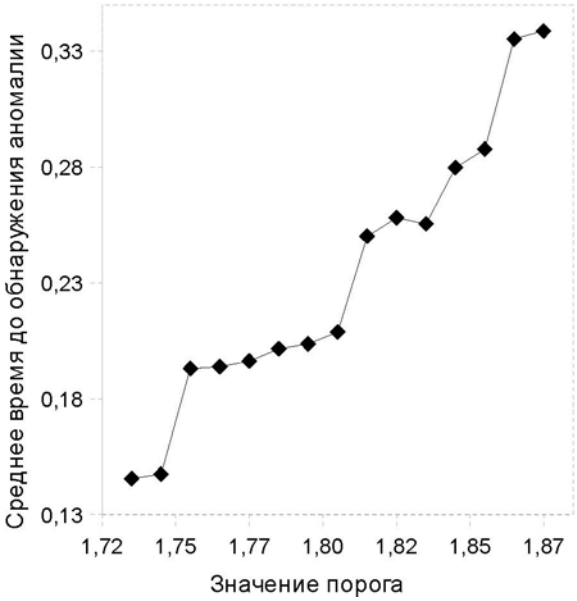

Рис. 3. Среднее время до обнаружения аномалии в зависимости от значения порога

4 Заключение

В работе рассматривается применение аномального подхода к задаче обнаружения полного несанкционированного копирования. Для апробации предлагаемого подхода была разработана программная система, строящая на основе исходных данных шаблон «нормального» поведения и классификатор. С помощью этой системы был проведен ряд вычислительных экспериментов, показавших применимость аномального подхода к задаче обнаружения полного несанкционированного копирования. Кроме того, было исследовано влияние значения порога на характеристики системы.

Итоговой целью работы является разработка системы защиты от несанкционированного полного копирования документов, основанной на подходе, представленном в данной работе и статьях [7, 8]. Такая система защиты может дополнить имеющиеся в ЭБ средства противодействия полному копированию.

Подход, представленный в статье, может быть использован для защиты от полного несанкционированного копирования не только текстовых документов, но и мультимедиа-контента (аудио- и видеозаписей).

Литература

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

[2] Wang J.-H., Chang H.-C., Hsiao J.-H. Protecting digital library collections with collaborative web image copy detection. – Digital Libraries: Universal and Ubiquitous Access to Information. – Berlin: Springer, 2008.

[3] Cox I.J., Miller M.L., Bloom J.A. Digital watermarking// The Morgan Kaufmann Series in Multimedia and Information Science. Series Editor: Edward Fox. – Morgan Kaufmann Publishers, USA. 2001.

[4] Koulouris A., Kapidakis S. Considerations on policies of university digital collections //Труды Шестой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», RCDL'2004, Пущино, Россия, 29 сентября – 1 октября 2004 г. – С. 159-168.

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

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

[7] Ивашко Е.Е. Построение системы защиты электронных библиотек от несанкционированного копирования документов //Труды Девятой Всерос. науч. конф. «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.

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

[9] Jha S., Tan K., Maxion R.A. Markov chains, classifiers and intrusion detection //Computer Security Foundations Workshop (CSFW), June 2001.

[10] Электронная библиотека Республики Карелия. – www.elibrary.karelia.ru.

Anomaly-based approach to detection of full unauthorized copying of documents

E.E. Ivashko, N.N. Nikitina

In the paper we present the recent results obtained in the process of developing the system of full unauthorized documents-copying detection for a digital library. The system is based on anomaly approach to intrusion detection and Markov chain method for modeling the pattern of "normal" behavior. The preceding results obtained at the earlier stages of work were presented in RCDL-2007 and RCDL-2009.

[i] Работа выполнена при финансовой поддержке РФФИ (проект 08-07-00085)