Опыт построения системы защиты электронных библиотек от несанкционированного копирования документов♣

© Ивашко Е. Е., Никитина Н. Н.

Карельский научный центр РАН
Институт прикладных математических исследований
{ivashko, nikitina}@krc.karelia.ru

Аннотация

В работе описаны результаты экспериментов, проведенных в рамках разработки системы защиты электронных библиотек от несанкционированного копирования документов. Статья представляет выводы, сделанные на основании практической проверки теоретических разработок, представленных на конференции RCDL-2007.

1 Введение

За последнее десятилетие электронные библиотеки (ЭБ) стали важной составляющей системы формирования и распространения научного знания. Свободный доступ к результатам исследований в различных областях является залогом дальнейшего развития науки. На создание и сопровождение коллекций электронных документов тратятся большие материальные и нематериальные ресурсы. При этом, зачастую дальнейшее развитие ЭБ ставится в прямую зависимость от посещаемости (популярности) ресурса.

Однако нередко полученные из ЭБ документы используются третьими лицами для получения прибыли в обход интересов правообладателей или для создания клона исходной ЭБ. Это делает актуальной задачу защиты ЭБ от полного несанкционированного копирования документов. Под полным несанкционированным копированием здесь и далее подразумевается получение электронных копий всех или большой части цифровых документов ЭБ без разрешения ее владельцев (правообладателей).

Основная идея, лежащая в основе исследования, представленного в данной работе и статье [1], заключается в следующем.

При использовании сервисов ЭБ пользователь решает актуальные для него задачи. При этом, обращаясь к различным электронным документам, он предполагает в рамках своих задач некоторую (возможно и несуществующую в действительности) субъективную семантическую связь между интересующими его документами. Например, студент при поиске материала для реферата по истории математики, может использовать биографии А. Пуанкаре и Г. Минковского, однако вряд ли будет обращаться к их математическим статьям и монографиям. Очевидно также, что ситуация, когда один и тот же пользователь интересуется одновременно узкоспециализированными темами из области физики, искусствоведения, генетики и др., является аномальной. Мы полагаем, что интерес к разнородным (семантически не связанным) документам является аномальным и может свидетельствовать о попытке копирования большой части разнородных документов в целях, связанных с нарушениями авторских прав (например, для создания «клона» исходной ЭБ).

Для обнаружения такого аномального поведения мы используем аномальный подход в обнаружении вторжений, основанный на предположении, что вторжение проявляется как отклонение от обычного («нормального») или ожидаемого поведения пользователя, и может быть обнаружено путем сравнения последовательности действий пользователя с некоторым заданным «шаблонным» поведением.

Здесь и далее полное несанкционированное копирование документов и вторжение трактуются в одном и том же смысле.

В данной работе представлены результаты экспериментов, проведенных в рамках разработки системы защиты электронных библиотек от полного несанкционированного копирования документов. Работа, описанная в статье, является продолжением исследований, представленных в рамках конференции RCDL в 2007 г. [1].

2 Описание модели

В этом разделе будут кратко описаны

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

аномальный подход в обнаружении вторжений и адаптированная технология обнаружения полного несанкционированного копирования документов. Более подробное формальное описание модели и связанных с ней вопросов можно найти в работе [1].

Аномальный подход в обнаружении вторжений основан на предположении, что вторжение проявляется как отклонение от обычного («нормального») или ожидаемого поведения пользователя, и может быть обнаружено путем сравнения последовательности действий пользователя с некоторым заданным «шаблонным» поведением.

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

1. построение «нормального» профиля поведения пользователя;
2. разработка классификатора, позволяющего отличить «нормальную» последовательность действий от аномальной;
3. определение граничных значений характеристик классификатора для снижения вероятности появления ошибок классификации;
4. обновление шаблонов «нормального» поведения.

Основой шаблона «нормального» поведения пользователя является Марковская цепь, построенная по записям поведения обычных пользователей ЭБ. Опишем кратко метод построения Марковской цепи.

Пусть имеется алфавит атомарных действий (например, список доступных документов ЭБ) ∑, множество всех конечных следов T* и тренировочный набор, составленный из заведомо нормальных следов $T_{tr} \in T^*$.

Расширим алфавит ∑специальным символом Ø. При построении МЦ задается параметр – «окно» размера w. Состояние в МЦ связано со следом длины w через алфавит ∑UØ, т. е. каждое состояние – набор из wсимволов алфавита ∑UØ. Переход – это пара (s, s'), определяющая в МЦ переход из состояния s в s'. Каждое состояние и переход также связаны со счетчиком количества переходов.

Операция shift(σ, x) сдвигает след σвлево 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 σи сдвигает σ на одну позицию влево, т.е. next(«abcd») возвращает a и обновляет след до «bcd».

Для каждого следа $\sigma \epsilon T_{tr}$, 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 sв 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.

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

Построенная по такому алгоритму МЦ представляет собой шаблон «нормального» поведения, который создается для каждого зарегистрированного в системе пользователя.

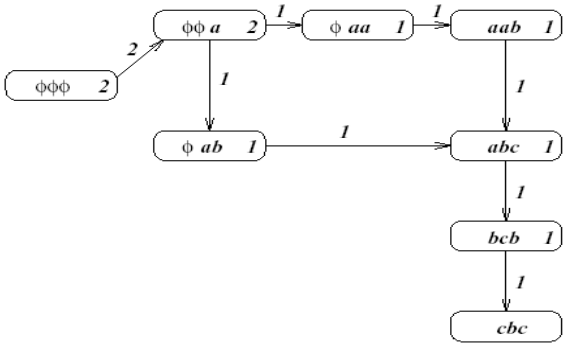

Рис. 1: Структура Марковской цепи

Целью работы, описываемой в данной статье, является проведение экспериментов и анализ их результатов, позволяющих на практике проверить эффективность модифицированного аномального метода обнаружения вторжений (представленного в [1]) применительно к обнаружению полного несанкционированного копирования документов ЭБ.

3 Описание экспериментов

Для проведения экспериментов была разработана программная система, выполняющая предварительную обработку файла исходных данных, построение профиля «нормального» поведения и проверку сессий работы пользователей на аномальность.

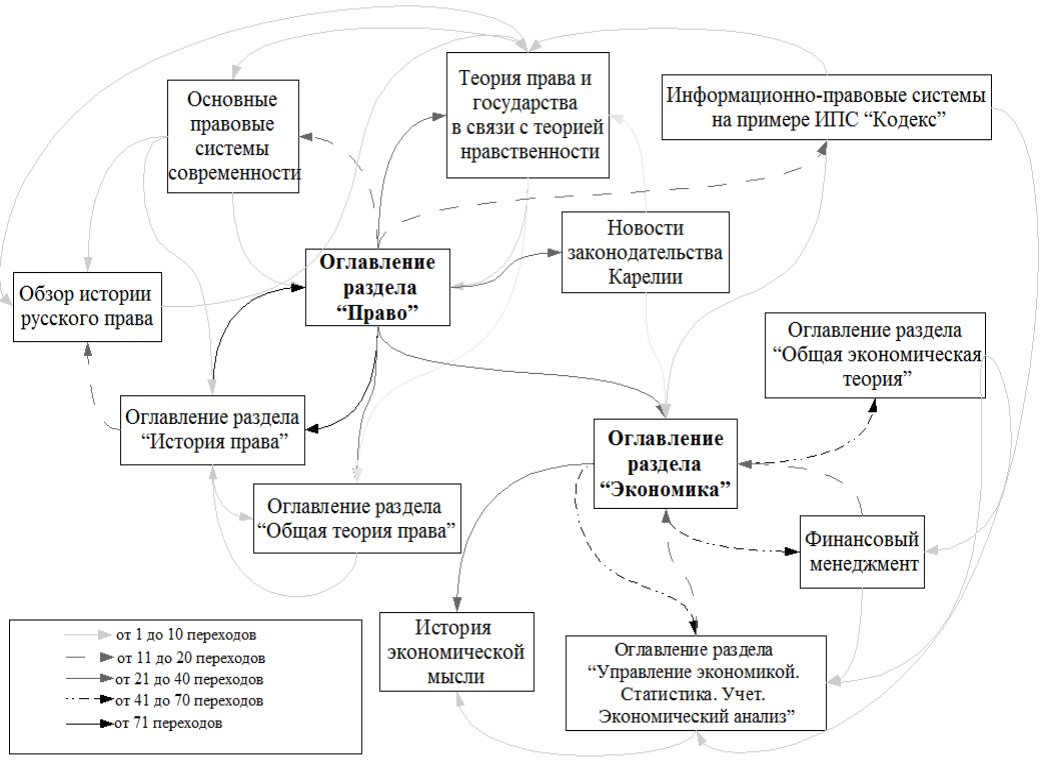

Рис. 2: Пример (фрагмент) профиля «нормального» поведения

3.1 Исходные данные

Исходными данными для проведения экспериментов послужили лог-файлы доступа к Электронной библиотеке Республики Карелия [2] 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.

Лог-файл обращений к цифровым документам записан в формате Common Log Format (CLF, [3]). Каждый запрос к серверу записан в отдельной строке, состоящей из полей, разделенных пробелами. При проведении экспериментов использовалась следующая информация, зафиксированная в лог-файле:

I— IP-адрес компьютера пользователя;

D — отметка времени (в CLF-формате);

R — строка запроса (содержит идентификатор запрашиваемого документа);

C — статус ответа сервера.

3.2 Преобразование исходных данных

Для построения шаблона «нормального» поведения была проведена предварительная обработка исходных данных. Несмотря на то, что в системе предусмотрена идентификация/аутентификация по паре имя пользователя/пароль, в лог-файле фиксируется лишь IP-адрес пользователя. Всего зафиксированы обращения с более 10000 различных IP-адресов. К сожалению, отсутствие в лог-файле записей об имени пользователя накладывает ограничения на возможности построения профиля «нормального» поведения, так как IP-адрес может являться адресом proxy-сервера, через который к документам ЭБ обращается одновременно несколько пользователей с различными интересами. Для того, чтобы такие обращения не влияли на итоговый результат, были отброшены записи, IP-адрес которых встречался в лог-файле более, чем в 4% случаев. Кроме того, были отброшены неинформативные сессии, содержащие менее 10 запросов к ЭБ.

Сессией работы пользователя считалась последовательность всех обращений к документам с конкретного IP-адреса. Всего в лог-файле содержится 4718 сессий работы пользователей.

3.3 Профиль «нормального» поведения

Согласно модели, для построения «нормального» профиля необходимы два набора данных: тренировочный (заведомо нормальные данные) и тестовый (для подбора оптимальных параметров). МЦ, являющаяся «нормальным» профилем поведения, была построена 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.

На рис. 2 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

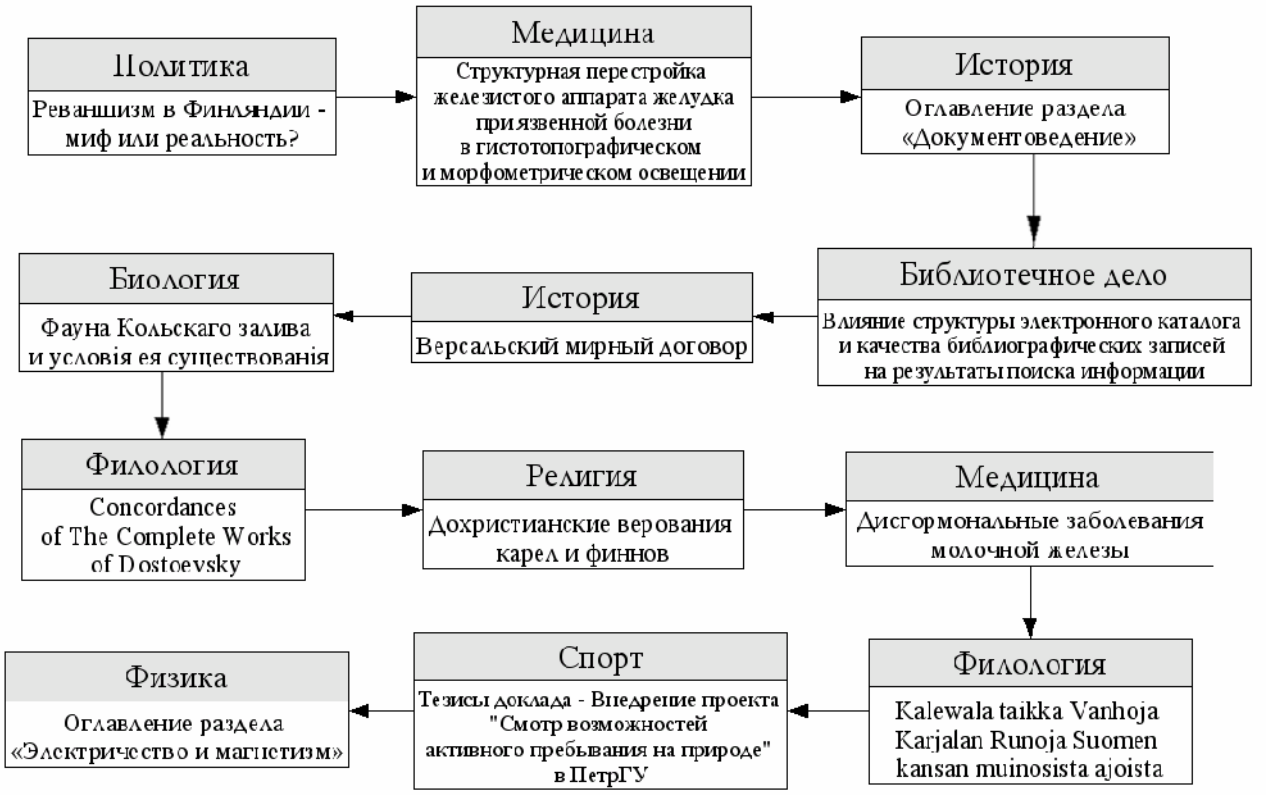

Рис. 3: Пример аномальной сессии работы пользователя

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.

3.4 Классификатор и аномальное поведение

Согласно модели аномального обнаружения вторжений [1], классификатор предназначен для определения значимых отличий проверяемой сессии работы пользователя от «нормального» поведения, представленного МЦ.

Одна из наиболее характерных аномальных сессий показана на рис. 3. В заголовке каждого запрошенного пользователем документа, указан раздел, в котором этот документ располагается в ЭБ.

Не вызывает сомнения, что такое разнообразие в выборе документов и разделов не является стандартным поведением пользователя. Выявление подобных аномальных сессий работы и является целью рассматриваемого в данной статье подхода.

4 Заключение

В работе представлены первые результаты ряда экспериментов, проведенных для проверки применимости и определения характеристик аномального подхода к защите ЭБ от 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.

Несмотря на некоторые ограничения, (связанные, в частности, с отсутствием в лог-файле информации о пользователях), можно сделать вывод о применимости аномального подхода в обнаружении вторжений к защите от 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.

При этом, однако, остается открытым вопрос, связанный с объемом данных по обращениям к документам, достаточным для построения полезных шаблонов нормального поведения. В работах, связанных с обнаружением вторжений на основе аномального подхода, как правило, указывается, что таких данных должно быть «достаточно много», однако какие-либо убедительные оценки (аналитические или эмпирические) отсутствуют.

В дальнейшем планируется сосредоточиться на подборе оптимальных параметров и определении следующих характеристик (согласно модели, представленной в [1]) подхода:

- размер окна при построении шаблона «нормального» поведения;
- количество ошибок классификации и среднее время до первого сообщения об аномальности сессии работы.

Итоговой целью работы является разработка

системы защиты от несанкционированного полного копирования документов, основанной на подходе, представленном в данной работе и статье [1], которая сможет дополнить имеющиеся в ЭБ средства защиты от копирования (например, ограничение числа документов, к которым может обратиться пользователь в единицу времени, и заключение договоров, гарантирующих права владельцев ЭБ). При этом, обнаружение аномального поведения в действиях пользователя может являться основанием для временного блокирования доступа пользователя к ресурсам ЭБ и проведения экспертизы.

Литература

[1] Ивашко Е. Е. Построение системы защиты электронных библиотек от несанкционированного копирования документов. //Труды Девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Переславль, Россия, 15-18 октября 2008 г. - Переславль-Залесский: изд-во «Университет города Переславля», 2007. С. 300-306.

[2] Электронная библиотека Республики Карелия. www.elibrary.karelia.ru.

[3] Описание формата Common Log Format. http://httpd.apache.org/docs/1.3/logs.html#common.

Some results of developing the unauthorized documents-copying protection system for digital libraries

E. Ivashko, N. Nikitina

In this article we consider results of the experiments made to check the workability of statistical anomaly detection algorithm to preserve digital libraries from unauthorized large-scale copying of documents. This work aims to check the theoretical model, introduced in RCDL-2007.

♣ Работа поддержана грантом РФФИ №08-07-00085а «Исследование технологических проблем создания и использования электронных коллекций научных информационных ресурсов»