Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

На правах рукописи

Трошин Сергей Владимирович

МОНИТОРИНГ РАБОТЫ ПОЛЬЗОВАТЕЛЕЙ КОРПОРАТИВНЫХ СЕТЕЙ

Специальность 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2010

Работа выполнена на кафедре автоматизации систем вычислительных комплексов факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова.

Научные руководители: доктор физико-математических наук, профессор **Машечкин Игорь Валерьевич**; кандидат физико-математических наук, доцент **Петровский Михаил Игоревич**

Официальные оппоненты: доктор физико-математических наук, член-корр. РАН **Воеводин Владимир Валентинович**; кандидат физико-математических наук, доцент **Головин Игорь Геннадьевич**

Ведущая организация: **Межведомственный суперкомпьютерный центр РАН**

Защита диссертации состоится « 23 » апреля 2010 г. в 11:00 на заседании диссертационного совета Д 501.001.44 в Московском государственном университете имени М. В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет вычислительной математики и кибернетики, ауд. 685.

С диссертацией можно ознакомиться в библиотеке факультета вычислительной математики и кибернетики МГУ имени М. В. Ломоносова. С текстом автореферата можно ознакомиться на официальном сайте факультета ВМК МГУ имени М. В. Ломоносова: http://cs.msu.su в разделе «Наука» – «Работа диссертационных советов» – «Д 501.001.44».

Автореферат разослан « 23 » марта 2010 г.

Ученый секретарь
диссертационного совета
профессор Н.П. Трифонов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Введение

Диссертация посвящена исследованию и разработке технологий построения специализированных систем мониторинга работы пользователей корпоративных сетей.

Предлагается новая технология построения систем мониторинга работы корпоративных пользователей, основанная на формировании и анализе моделей поведения отдельных пользователей и пользовательских групп. Данная технология позволяет создавать специализированные системы мониторинга, предназначенные для решения таких актуальных задач как:

- обнаружение действий пользователей, которые могут предшествовать внутренним вторжениям[1] (далее «раннее обнаружение внутренних вторжений»);
- обнаружение некомпетентных действий пользователей, а также случаев нецелевого использования вычислительных и информационных ресурсов корпоративной сети.

Общей характеристикой данных задач является то, что при их решении затруднено использование экспертных знаний, описывающих сценарии «интересных» для аналитика действий пользователей. Для внутренних вторжений практически отсутствуют сценарии, описывающие последовательность действий, определяющих подготовку внутреннего вторжения. Для нецелевого и непрофессионального использования не могут быть созданы однозначные правила, так как такие действия могут трактоваться каждой организацией по-своему и зависеть от контекста работы пользователей. Таким образом, традиционные подходы, применяемые в большинстве систем информационной безопасности, основанные на использовании сигнатурных баз данных, описывающих атаки или нецелевые действия, не могут обеспечить полное решение задачи.

Актуальной является разработка технологии, позволяющей на основе сбора, консолидации и централизованного анализа данных формировать модели корректного поведения пользователей (или групп пользователей), затем, путем сопоставления текущих действий пользователей с построенными моделями, решать обозначенные выше задачи. Выявляя аномальные относительно моделей события, можно определять действия пользователей, которые могут являться фактами нецелевого или непрофессионального использования, подготовкой внутренних вторжений или иными отклонениями в работе, требующими внимания аналитика. Для проведения такого анализа аналитик должен выбрать «эталонные» данные

[1]Под внутренним вторжением понимаются действия легального пользователя, направленные на нарушение целостности, конфиденциальности или доступности данных корпоративной сети.

для построения моделей поведения и данные, которые необходимо оценить на основе построенных моделей. Особенностью такого подхода является проведение анализа в отложенном режиме, т.е. по требованию аналитика на основе заданных им параметров. Данная схема исключает необходимость поиска и формирования экспертами сигнатур и правил, характеризующих обнаруживаемые действия.

Важной характеристикой систем мониторинга является тип данных, который используется для анализа. Большинство существующих систем информационной безопасности построено на основе анализа содержательной (контентной) информации. Для поиска вторжений, могут составляться наборы правил, объединяющих ключевые слова и признаки документов. Эти правила используются для контроля содержимого электронных документов. Такой подход имеет целый ряд недостатков, среди которых главным является то, что анализ контентной информации потенциально снижает уровень защищенности корпоративных данных в сети. Это происходит по следующим причинам:

- Эксперты, настраивающие систему безопасности, получают доступ к конфиденциальной информации.
- Вся информация проходит через систему, что может приводить к возникновению новых потенциальных уязвимостей.

Актуальным является исключение использования контентной информации при проведении мониторинга. Требуется разработка технологии, основанной на использовании журналируемой информации, т.е. информации, не являющейся содержимым файлов или иных документов, а являющейся совокупностью описания событий работы пользователей или программ. Такие события могут фиксироваться операционной системой, прикладными программами или специализированными системами.

Основной характеристикой новой технологии является возможность создания систем мониторинга, учитывающих особенности корпоративной сети: размеры сети и состав наблюдаемых систем, цели мониторинга и набор доступной информации. В связи с этим, актуальным является разработка унифицированных моделей и методов, позволяющих производить сбор, консолидацию и анализ журналированных событий из различных источников и наблюдаемых систем, с учетом особенностей и ограничений использования сети передачи данных и других компонентов корпоративной сети.

Цель диссертационной работы

Целью диссертации является исследование методов, алгоритмов и подходов проведения мониторинга и разработка, на основе результатов исследования, новой технологии построения специализированных систем мониторинга работы пользователей с ресурсами корпоративной сети. Технология должна основываться на использовании журналируемой информации и реализовывать моделирование поведения и поиск аномалий в

работе пользователей. Построенные на основе технологии системы должны позволять накапливать информацию, описывающую параметры работы пользователей, и проводить отложенный анализ с целью решения ряда специфических задач, таких как раннее обнаружение внутренних вторжений, выявление нецелевого и непрофессионального использования ресурсов корпоративной сети.

Научная новизна работы

Предложенная в работе технология построена на новых принципах раннего обнаружения внутренних вторжений, а так же нецелевого и непрофессионального использования ресурсов корпоративной сети. Предложенные принципы основаны на моделировании поведения пользователей, поиске аномалий в действиях пользователей и изменений состава используемых ресурсов и параметров использования ресурсов.

Практическая значимость

Предложенная в работе технология может быть использована для построения систем мониторинга работы корпоративных пользователей, позволяющих решать такие актуальные задачи как раннее обнаружение внутренних вторжений, обнаружение некорректного и нецелевого использования корпоративных ресурсов, анализ статистики работы пользователей. В свою очередь это позволяет минимизировать утечки конфиденциальной информации, а так же оптимизировать рабочий процесс и использование ресурсов организации.

Методы исследования

В работе использовались алгоритмы интеллектуального анализа данных (data mining) и технология оперативной аналитической обработки OLAP. При проектировании и разработке экспериментальной системы мониторинга применялся объектно-ориентированный подход и мультиагентная архитектура.

Апробация работы и публикации

Основные результаты диссертации опубликованы в девяти печатных работах [1-9]. Статья [9] опубликована в издании, рекомендованном ВАК для публикации результатов кандидатских диссертаций. Основные положения работы докладывались на следующих конференциях и семинарах:

1. XIII Международная конференция студентов, аспирантов и молодых учёных «ЛОМОНОСОВ-2006», МГУ, Москва, 2006 г.
2. Конференция «Ломоносовские чтения 2006», МГУ, Москва, 2006 г.
3. Конференция «Тихоновские чтения», МГУ, Москва, 2006 г.
4. First Spring Young Researches’ Colloquium on Software Engineering (SYRCoSE’2007), Moscow, Russia, 2007.

5. Вторая международная конференция «Системный анализ и информационные технологии» САИТ-2007, Обнинск, Россия, 2007 г.
6. 13-я Всероссийская конференция «Математические методы распознавания образов», Зеленогорск, Россия, 2007 г.
7. 6-я международная конференция по программированию «УкрПРОГ'2008», Киев, Украина, 2008 г.

Структура и объем работы

Диссертация состоит из введения, трех глав, заключения, списка литературы и приложений. Объем диссертации с приложениями составляет 170 страниц. Список литературы включает 88 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** сформулированы цели и задачи работы. Обоснована ее актуальность.

Первая глава посвящена исследованию и сравнительному анализу существующих методов и подходов к построению систем мониторинга. В главе сформулированы и обоснованы требования к системам мониторинга, основанным на сборе, консолидации и анализе журналируемой информации. Основными требованиями являются: расширяемость набора источников данных; масштабируемость системы и возможность работы в крупных сетях; проведение фильтрации, нормализации, агрегации и корреляции собираемых данных; оперативная аналитическая обработка уже накопленных данных, поиск аномальных событий и отклонений в работе пользователей. В результате проведенного исследования отобран ряд широко используемых систем управления информацией и событиями безопасности (SIEM – Security Information and Event Management), а так же систем мониторинга работы персонала. С целью уточнения направлений исследования проведен сравнительный анализ их возможностей, используемых методов и подходов. Основными анализируемыми параметрами являлись: методы сбора журналируемых событий, методы долговременного централизованного хранения данных, методы построения статистических ответов и методы поиска «значимых» событий. В результате анализа показано, что на сегодняшний день не существует систем мониторинга, удовлетворяющих всем поставленным требованиям и решающих поставленные задачи.

Сформулированы основные системные и архитектурные концепции построения систем мониторинга работы корпоративных пользователей, а так же направления исследования. Мониторинг должен быть основан на моделировании поведения пользователей и поиске аномалий в работе пользователей. При этом модель поведения и методы анализа должны позволять обнаруживать как отдельные аномальные действия пользователей, так и изменения состава используемых ресурсов и параметров их использования, таких как количество, частота, продолжительность и т.п.

Системы мониторинга должны использовать журналируемую информацию и позволять анализировать работу пользователей с различными ресурсами организации. Системы должны обеспечивать производительность, достаточную для работы в современных корпоративных сетях, минимизировать дополнительную нагрузку на наблюдаемые компьютеры и корпоративную сеть, а так же обеспечивать защищенность данных от аварийных ситуаций и противодействий со стороны пользователей.

Во **второй главе** описывается разработанная в рамках диссертационной работы модель мониторинга, основанная на формализации сбора и обработки данных, описывающих работу пользователей с ресурсами. Построение модели предполагает выбор формата исходных данных и разработки методов их предобработки, формальное описание представления моделей поведения пользователей, разработку методов построения моделей поведения и методов поиска аномалий на их основе. Предполагается, что данная модель должна служить основой для разработки технологии мониторинга работы пользователей с ресурсами корпоративной сети.

Выбор формата исходных данных и разработка методов их предобработки. В качестве исходных данных мониторинга в работе предложено использовать системные и прикладные журналы. Журналируемая информация содержит описание отдельных событий, например: «запуск пользователем User1 процесса telnet.exe» или «передача пакета данных процессом telnet.exe». Под событием понимается набор $event = \langle type, time, attr_1, ..., attr_n \rangle$, где $type$ – тип события, $time$ – время события, а $attr_1, ..., attr_n$ – набор типизированных атрибутов события, при этом множество атрибутов не всегда указывает на пользователя и на ресурс. Для анализа параметров работы пользователей с ресурсами представляют интерес не только отдельные события, но и их множества и последовательности, описывающие совокупное действие, например, формирование процесса и суммарный объем данных переданный этим процессом. Таким образом, требуется проведение агрегации журналируемых событий в новые более информативные события, описывающие законченную работу пользователей с ресурсами.

Для описания агрегированных событий вводится понятие факта активности $fact = \langle type, user, computer, time, a_1, ..., a_m \rangle$, где $type$ – тип ресурса (факта активности), $user, computer, time$ – атрибуты, описывающие пользователя, компьютер и время, $a_1, ..., a_m$ – набор именованных свойств (атрибутов) факта. Числовые атрибуты описывают параметры работы с ресурсом, а текстовые – ресурс и действие над ресурсом.

Предложен метод построения фактов активности, основанный на выделении из потока журналируемых событий цепочек событий,

относящихся к одному факту. Входными данными метода являются упорядоченный по времени поток журналируемых событий $EV = \{event_1, event_2, ...\}$ и формализованные сценарии построения фактов. Результатом работы метода является множество фактов $F = \{ fact_1, fact_2, ...\}$.

Сценарий построения факта активности – это автомат, дополненный условиями и действиями $S = \langle S, Type, Filter, t, Env, fact \rangle$, где

- S – конечное множество состояний, среди которых выделяется начальное и конечное состояние;
- $Type$ – конечное множество типов агрегируемых событий;
- $Filter$ – конечное множество условий на значения атрибутов событий. Каждое условие задается в виде $attr_j = v_j$, где v_j – требуемое значение атрибута;
- t – функция переходов между состояниями $t : S \times \langle Type, Filter \rangle \to S$, переход однозначно определяется текущим состоянием и типом пришедшего события, но осуществляется, только при выполнении соответствующих условий на атрибуты события, заданных в Filter. В случае невозможности перехода событие игнорируется, а состояние не изменяется;
- Env – множество переменных окружения сценария. Переменные используются для хранения промежуточных данных агрегации и служебных переменных;
- $fact$ – формируемый факт активности с необходимыми атрибутами, $\{ fact.a \}$ – множество атрибутов факта.

Каждому состоянию $s \in S$ ставятся в соответствие три множества функций, вычисляемых после перехода в это состояние: $s = \langle \{ ffunc \}, \{efunc\}, \{tfunc\} \rangle$, где $ffunc$, $efunc$ и $tfunc$ – функции, определенные над переменными окружения, атрибутами факта и атрибутами пришедшего события: $ffunc : Env \times \{ fact.a \} \times \{event.attr\} \to \{ fact.a \}$ изменяет значение атрибута формируемого факта; $efunc : Env \times \{ fact.a \} \times \{event.attr\} \to Env$ изменяет значение переменной окружения; $tfunc : Env \times \{ fact.a \} \times \{event.attr\} \to Filter$ изменяет условие на значение атрибута события (функцию перехода). Факт активности считается сформированным, когда автомат приходит в заключительное состояние.

Сценарий задается экспертом в виде регулярной грамматики. Множество терминальных символов соответствует типам событий $Type$. Множество нетерминальных символов – множеству состояний S. Дополнительно эксперт указывает условия на значения атрибутов события для каждого правила вывода (перехода), а так же функции $\{ ffunc \}, \{efunc\}, \{tfunc\}$ для каждого нетерминального символа.

Множество построенных фактов активности используется для построения модели поведения пользователей. Модель поведения предложено представлять в виде пары $M = \langle \{R\}, Prof \rangle$, где $\{R\}$ – множество шаблонов поведения пользователей (правил, описывающих связь значений атрибутов отдельных фактов активности), *Prof* – профиль использования ресурсов (статистические параметры работы, вычисленные на основе распределения значений свойств фактов активности) (рис. 1).

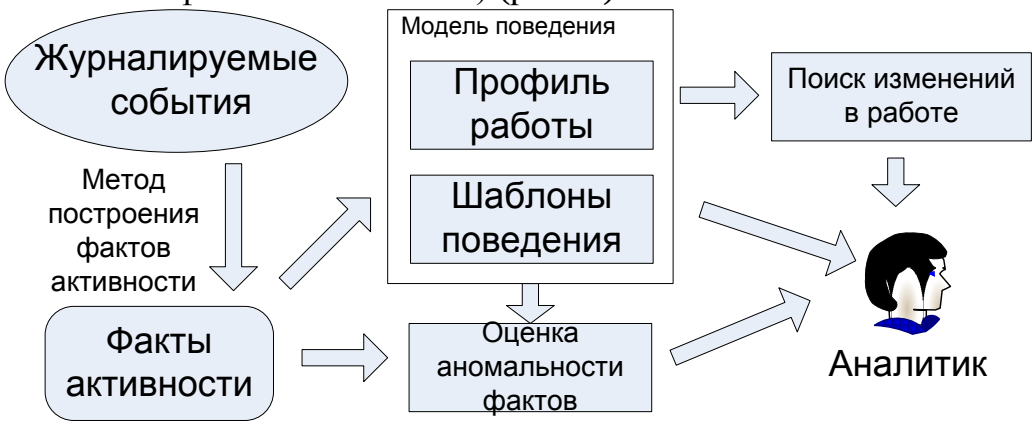

Рис. 1. Концептуальная схема использования модели.

Основные шаги предложенной схемы мониторинга следующие:

1. На основе выбранного аналитиком подмножества фактов активности строиться модель поведения пользователей.
2. Шаблоны поведения используются для поиска отдельных аномальных фактов активности.
3. Профили используются для поиска изменений параметров использования ресурсов.
4. Найденные изменения и аномалии могут визуализироваться аналитику.

Далее в главе рассматривается построение шаблонов поведения и их использование для поиска аномалий, а так же построение профилей работы и их использование для поиска изменений в статистике работы.

Формальное описание представления шаблонов поведения. Шаблоны поведения предложено описывать в виде ассоциативных правил над атрибутами фактов активности. С целью построения шаблонов поведения аналитиком выделяется обучающий набор однотипных фактов активности, обычно, описывающих типовую работу определенных пользователей за определенное время. Затем выбирается подмножество атрибутов фактов, для которых будут строиться шаблоны. Перед построением правил проводится дискретизация числовых атрибутов с помощью алгоритма expectation maximization(EM)[2]. Для построения ассоциативных правил используется алгоритм Apriori[2].

Каждое ассоциативное правило *R* определено как импликация:

[2] А.А. Барсегян, М.С. Куприянов // Методы и модели анализа данных: OLAP и Data Mining. Санкт-Петербург, 2004, сс. 129-147.

$$R = "A_1(a_1),...,A_n(a_n) \Rightarrow B_1(a_1),...,B_m(a_m)" \qquad [c, s]$$

где: $A_l(a_l), B_l(a_l)$ – условия на значения *l-го* атрибута факта активности. Для дискретных атрибутов используется условие «равно», для числовых «принадлежит интервалу» (например, process=«telnet.exe» или duration=«12-653»), s – поддержка правила. Правило $A_1(a_1),...,A_n(a_n) \Rightarrow B_1(a_1),...,B_m(a_m)$ имеет поддержку s, если s процентов фактов активности обучающего набора содержат атрибуты, удовлетворяющие условиям $A_1(a_1),...,A_n(a_n) \cup B_1(a_1),...,B_m(a_m)$.

c – достоверность правила определенная как:

$$c = \text{confidence}(R) = \frac{s\left(A_1(a_1),...,A_n(a_n) \cup B_1(a_1),...,B_m(a_m)\right)}{s\left(A_1(a_1),...,A_n(a_n)\right)}$$

Ассоциативные правила могут быть построены по произвольному фиксированному набору атрибутов фактов активности, и описывать зависимости параметров работы пользователей, например: «пользователь = User1 и процесс = telnet.exe => адрес = 158.250.19.39 с достоверностью 0.9».

Метод поиска аномалий. Шаблоны предложено использовать для поиска отдельных аномалий в работе пользователей, что соответствует обнаружению отдельных фактов активности пользователя, существенно отличающихся от шаблонов поведения, описанных моделью поведения. Для этого аналитиком выбирается набор фактов активности, описывающий работу пользователей, которую требуется проверить на наличие аномалий.

Предложенный метод оценки степени аномальности факта активности $fact = \left(a_1 = v_1,...,a_n = v_n\right)$ основывается на использовании множества построенных ассоциативных правил для прогнозирования значений одного атрибута факта по значениям остальных атрибутов: вычисляется функция оценки ожидаемости значения атрибута $P(a_i = v \mid a_1 = v_1,...,a_{i-1} = v_{i-1}, a_{i+1} = v_{i+1},...,a_n = v_n)$, определенная как

$$\max_j \text{confidence}(R_j(a_1 = v_1,...,a_{i-1} = v_{i-1}, a_i = v, a_{i+1} = v_{i+1},...,a_n = v_n))$$

С целью вычисления такой функции для выбранного факта и его атрибутов $fact = \left(a_1 = v_1,...,a_n = v_n\right)$ ищутся все правила, где условия на атрибуты левой части правила удовлетворяют значениям $a_1 = v_1,...,a_{i-1} = v_{i-1}, a_{i+1} = v_{i+1},...,a_n = v_n$. Среди таких правил выбираются те, где условия правой части соответствуют значению атрибута $a_i = v$. Рассчитывается совокупная достоверность по системе таких правил как максимальная достоверность среди правил.

Затем вычисляется степень нормальности значения атрибута $n\left(a_i = v \mid a_1 = v_1,...,a_{i-1} = v_{i-1}, a_{i+1} = v_{i+1},...,a_n = v_n\right)$, как отношение ожидаемости наблюдаемого значения v к максимальной ожидаемости значений атрибута:

$$\frac{P(a_i = v \mid a_1 = v_1, ..., a_{i-1} = v_{i-1}, a_{i+1} = v_{i+1}, ..., a_n = v_n)}{\max\limits_{dom(a_i)} P(a_i \mid a_1 = v_1, ..., a_{i-1} = v_{i-1}, a_{i+1} = v_{i+1}, ..., a_n = v_n)}$$

Если значение атрибута a_i совпадает с наиболее ожидаемым, т.е. тем, которое прогнозируется на основе ассоциативных правил, то нормальность такого атрибута равна 1. Если же такое значение ранее вообще не встречалось, то нормальность атрибута будет равна нулю, что соответствует «аномальному» значению. В остальных случаях значение нормальности будет меняться от 0 до 1.

В случае если для значения атрибута не найдено ассоциативных правил (значение не зависит от других атрибутов), то нормальность атрибута вычисляется следующим образом:

$$n\left(a_i = v \mid a_1 = v_1, ..., a_{i-1} = v_{i-1}, a_{i+1} = v_{i+1}, ..., a_n = v_n\right) = \frac{s(a_i = v)}{\max\limits_{dom(a_i)} s(a_i)},$$

где s – поддержка значения атрибута, т.е. маргинальная вероятность появления указанного значения для данного атрибута (может рассматриваться как вырожденное ассоциативное правило, где условие состоит из одного атрибута, и нет правой части).

Аномальность факта вычисляется на основе степеней нормальности его атрибутов:

$$anom(fact) = 1 - \prod_i n\left(a_i = v \mid a_1 = v_1, ..., a_{i-1} = v_{i-1}, a_{i+1} = v_{i+1}, ..., a_n = v_n\right)$$

Представленный подход позволяет интерпретировать найденные аномалии, т.е. определять какие комбинации из значений атрибутов, описывающих ресурс, пользователя или параметры работы привели к аномальности факта.

Факты активности описывают отдельные действия пользователей, но не учитывают статистику таких действий. В частности, не описывается интенсивность обращений к ресурсу, распределение обращений во времени и суммарные данные использования. Статистика использования ресурсов необходима для поиска изменений состава используемых ресурсов и параметров их использования.

Формальное описание представления профилей работы. Для расчета и хранения статистики работы пользователей предложена специальная структура (профиль), содержащая агрегированные параметры использования ресурсов в зависимости от пользователя, действий с ресурсом и временных интервалов. Примером таких агрегированных параметров может быть зависимость суммы переданных байт от процесса, пользователя и времени: traffic(‘User1’, ‘telnet.exe’, ‘Март’)=431, traffic(‘User2’, ‘telnet.exe’, ‘Март’)=0, traffic(‘User1’, ‘ping.exe’, ‘Март’)=1530, traffic(‘User2’, ‘ping.exe’, ‘Март’)=14252. Профиль описывается парой $Prof = \langle D, M \rangle$, где

- D – конечное множество дискретных атрибутов факта активности.
- M – конечное множество агрегированных значений числовых атрибутов фактов активности.

С целью построения профиля для каждой требуемой функции агрегирования A задается отображение $A: D_1 \times ... \times D_n \times \{fact.a\} \to M, n \le |D|$. Для этого предложено использовать технологию оперативной аналитической обработки данных OLAP. Результатом применения технологии является OLAP-куб – т.е. набор представлений. Пример куба для фактов передачи данных различными пользователями с использованием различных процессов приведен на рис. 2. Названия узлов обозначают атрибуты факта, по которым в представлении имеются данные, по значениям «ALL» производится агрегация.

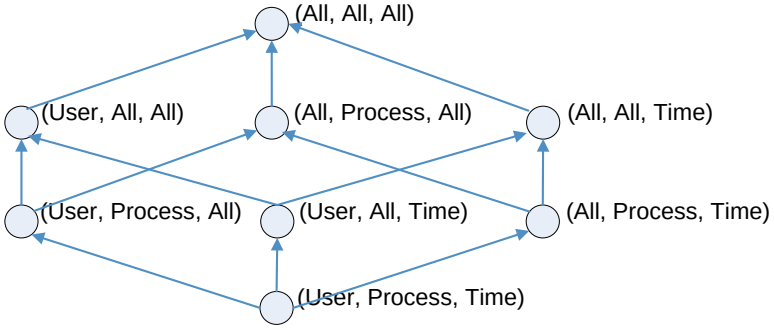

Рис. 2. Пример представления куба.

Для вычисления агрегированных параметров фиксируется набор дискретных атрибутов и выбирается соответствующее представление в кубе. Например, для получения суммарного объема переданных пользователем $User1$ данных в зависимости от процесса, на основе представления «User, Process, All» может быть построено отображение $D_{process} \to SumBytesSend$, при условии что $D_{user} = 'User1', A = 'sum', I = 'bytesSend'$.

Построенные отображения описывают состав используемых ресурсов и распределение параметров их использования. Отображения рассматриваются как гистограммы, которые могут быть визуализированы эксперту, а так же применяются для поиска изменений в работе. Предложенный метод поиска изменений основан на сравнении двух гистограмм от одинаковых наборов атрибутов и рассматривается отдельно для случая изменения работы пользователя во времени и для сравнения работы пользователей и групп.

Метод поиска изменений в работе. С целью обнаружения изменения работы пользователя во времени сравниваются две гистограммы параметров использования ресурсов, определенных над одним множеством дискретных атрибутов $D_1, ..., D_m, m < n$, но для различных временных интервалов: $time_1$ («эталонный») и $time_2$ («проверяемый»). Для сравнения гистограмм

используются критерий согласия Пирсона. В случае непринятия гипотезы о соответствии гистограммы за временной интервал $time_2$, гистограмме за интервал $time_1$, делается вывод об изменении рассматриваемого параметра работы пользователя с ресурсами, описываемыми атрибутами $D_1,...,D_m, m \le n$.

Для сравнения работы группы пользователей предложен метод, основанный на вычислении меры сходства гистограмм для двух пользователей. Для этого используется симметричное расстояние хи-квадрат:

$$d = \frac{1}{2} \sum_{(d_1,...,d_m) \in D_1 \times ... \times D_m} \frac{\left(func_{user2}'(d_1,...,d_m) - func_{user1}'(d_1,...,d_m) \right)^2}{func_{user2}'(d_1,...,d_m) + func_{user1}'(d_1,...,d_m)},$$

где $func'_{user1}$ и $func'_{user2}$ функции распределения, построенные по гистограммам, описывающим работу пользователей $user_1$ и $user_2$, и определенным над множеством $D_1,...,D_m, m < n$. Вычисленные меры используются для иерархической кластеризации пользователей. Результаты кластеризации могут быть отображены в виде дендрограммы, что позволяет находить группы похожих пользователей, и пользователей, отличающихся от других пользователей, или находить различные группы пользователей.

Предложенная модель мониторинга, позволяет реализовать концепцию раннего обнаружения с помощью применения методов интеллектуального анализа данных для поиска аномальных действий пользователей, технологии OLAP для получения агрегированных показателей работы пользователей и сравнения статистики с целью поиска изменений в работе. Найденные изменения и аномалии передаются эксперту и могут трактоваться как нарушения, либо быть сигналом к усилению контроля над определенными параметрами работы определенных пользователей.

Третья глава посвящена исследованию методов и подходов к организации сбора, консолидации и централизованного хранению данных, полученных на основе журналируемой информации, а так же разработке технологии построения систем мониторинга, реализующей предложенную модель.

Основными требованиями к системам мониторинга являются:

1. масштабируемость систем для работы в крупных сетях;
2. расширяемость набора собираемых исходных данных;
3. минимизация дополнительной нагрузки на процессор, память и диск наблюдаемых систем, а так же на сеть передачи;
4. долговременное хранение собранных фактов активности;
5. обеспечение защищенности собираемых и обрабатываемых данных от порчи, кражи, подмены и потери в результате сбоев в работе компонентов сети.

Предложено системное решение, основанное на мультиагентной архитектуре, которая является масштабируемой за счет тиражирования

агентов и расширяемой за счет создания агентов для новых типов наблюдаемых систем (рис. 3).

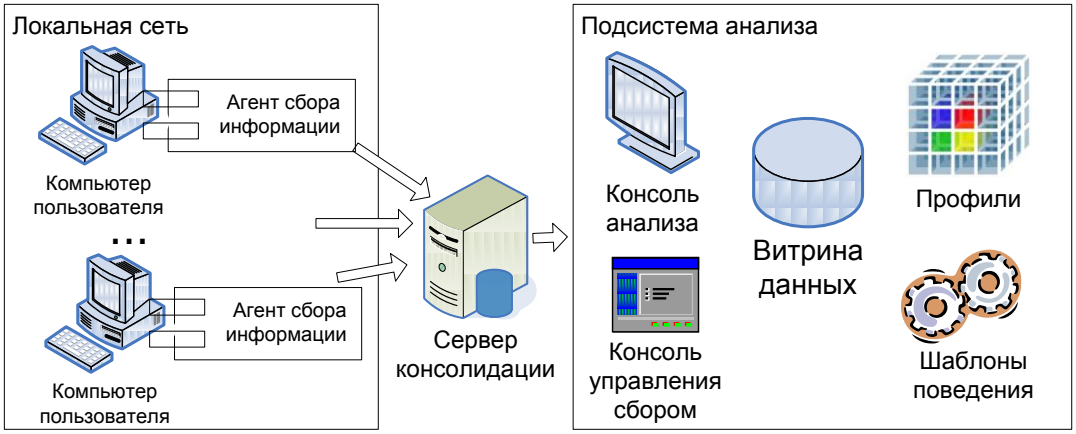

Рис. 3. Общая архитектура системы.

Агент осуществляет чтение журналируемых данных, формирует факты активности и передает их на сервер консолидации, который размещает факты в хранилище. Подсистема анализа запрашивает с сервера выбранное аналитиком подмножество фактов активности в витрину данных и проводит их анализ. Архитектура допускает использование нескольких подсистем анализа данных, нескольких серверов консолидации и проведение иерархической консолидации.

Предложен модульный принцип построения агента сбора, основанный на использовании системно-независимых модулей (отмечено штрихом на рис. 4) и модулей, требующих разработки для каждого типа системы.

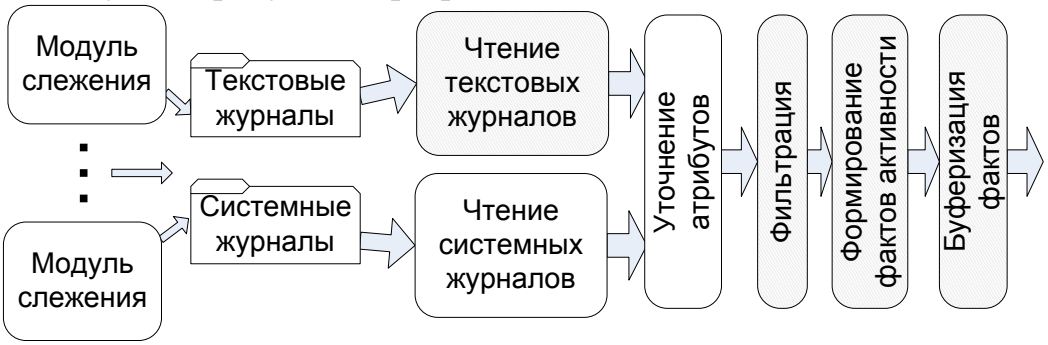

Рисунок 4. Схема обработки данных в подсистеме сбора.

Расширяемость набора исходных данных достигается за счет настраиваемого модуля чтения текстовых журналов и возможности добавления модулей слежения за работой пользователей, каждый из которых сохраняет собранные параметры в журналах. Это обеспечивает защищенность данных, а так же унифицирует работу с различными модулями за счет чтения данных из одного или нескольких журналов.

Модуль уточнения атрибутов реализует преобразования, которые возможны только на наблюдаемой системе, например: определение адресов по IP-адресам, восстановление имен пользователей по их системным идентификаторам и пр. Так же модуль реализует определение процессов, в

рамках которых происходят действия пользователей, для этого поддерживается список запущенных в системе процессов.

Агент буферизирует построенные факты активности и передает данные на сервер согласно стратегиям: фиксированный объем, фиксированное количество фактов или через равные временные промежутки. Буфер защищен от сбоев в работе наблюдаемой системы и от противодействий со стороны пользователей за счет избыточного копирования и средств разграничения доступа файловой системы. Передаваемые данные шифруются с помощью криптографического протокола SSL.

С целью построения сервера консолидации и хранилища данных оценена требуемая производительность импорта и экспорта фактов активности, а так же сформулированы ограничения на использование дискового пространства. Проведен сравнительный анализ времени добавления и извлечения данных и эффективности использования дискового пространства традиционными решениями, такими как текстовые форматы и реляционные СУБД (Oracle 10g и MS SQL 2008). В результате анализа показано, что рассмотренные средства не удовлетворяют поставленным требованиям для сетей насчитывающих более 2 тысяч компьютеров.

Для хранения фактов активности разработано специализированное хранилище, основанное на использовании файловой системы, учитывающее особенности данных и основных операций (рис. 5).

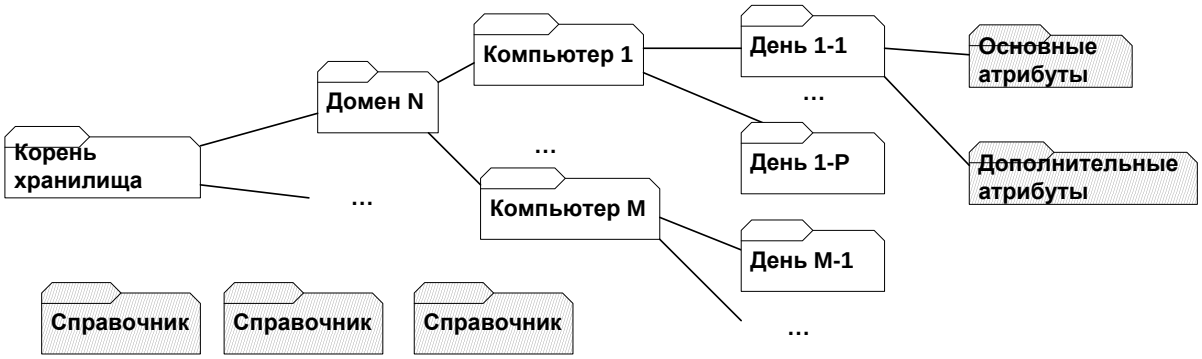

Рис. 5. Структура файлов хранилища.

Запись о факте активности разделяется между двумя файлами данных и файлами-справочниками. Два файла используются для хранения основных (тип факта, пользователь, компьютер, время) и дополнительных атрибутов, соответственно, что позволяет производить поиск по основным атрибутам, не считывая полную запись о факте. Значения текстовых атрибутов, имена атрибутов, имена пользователей и компьютеров хранятся в трех справочниках, соответственно. Это позволило уменьшить объем хранилища, без ущерба производительности, за счет хранения часто используемых данных в небольших справочниках. Данные хранилища организованы в древовидную структуру, что позволяет быстро получать «срезы» по компьютерам и дням, а так же решает проблемы копирования и запроса части хранилища за счет обращения только к необходимым файлам.

Для обеспечения надежности хранилища реализуется механизм транзакций для добавления данных, основанный на реализации собственных механизмов контрольных точек файлов и избыточного хранения для справочников. Требование защищенности данных удовлетворяется за счет средств разграничения доступа ОС.

Построение профилей основано на применении технологии OLAP, для чего факты активности отображаются на специальную структуру витрины данных типа «снежинка». В центре структуры находятся таблицы фактов, которые содержат значения числовых атрибутов и внешние ключи на таблицы измерений, где хранятся текстовые атрибуты. В таблицы измерений отображаются записи справочников. С целью повышения наглядности отчетов отдельно строятся таблицы с иерархическими измерениями для основных атрибутов «пользователь» и «компьютер», а так же для часто встречаемых атрибутов «имя процесса» и «имя файла». Построение OLAP-кубов по витрине данных осуществляется средствами реализации технологии OLAP продукта Microsoft Analysis Services 2005.

Шаблоны поведения пользователей строятся на основе фактов активности из витрины данных. Для построения шаблонов используется алгоритм Apriori. Методы оценки степени аномальности фактов активности реализованы в работе на языке Си++, что позволило добиться необходимой производительности.

Для заполнения витрины данных предложено использовать механизм объемной вставки (bulk insert). Считанные из файлов хранилища факты активности, а так же необходимые значения справочников отображаются во временные текстовые файлы, которые затем без перерасчета идентификаторов и внешних ключей помещаются в базу.

Заполнение витрины данных, построение моделей поведения и поиск аномалий производится в фоновом режиме, что позволяет обеспечить требуемую производительность анализа за счет использования вычислительных ресурсов в то время, пока аналитик просматривает уже готовые материалы.

В рамках апробации технологии была построена экспериментальная система мониторинга, собирающая следующие факты активности: сеансы работы пользователей; запуск процессов; работа с клавиатурой и мышью; сетевые TCP соединения; работа с файловой системой; изменение состава программ и аппаратной конфигурации.

Производительность системы оценивалась на компьютере со следующими характеристиками: Intel Xeon 2,83 ГГц (2 шт.); 8 Гб; Windows Server 2003 R2 64bit. Скорость приема и импорта данных от агентов составила 72 миллиона фактов активности в сутки, т.е. для работы в сетях, насчитывающих 10 тысяч компьютеров, достаточно одного сервера консолидации.

Оценка производительности подсистемы анализа была проведена для основных сценариев использования, а именно: запрос выбранных аналитиком данных с сервера консолидации и заполнение витрины данных, построение модели поведения (профили работы и шаблоны поведения) пользователей по выбранным фактам, оценка степени аномальности подмножества фактов на основе модели поведения (табл. 1).

Сценарии	100 тыс. фактов	250 тыс. фактов	500 тыс. фактов	1 млн. фактов
Заполнение витрины	28 сек	1 мин	2 мин	4 мин
Построение профилей	5 сек	7 сек	12 сек	20 сек
Построение шаблонов	8 сек	15сек	31 сек	64 сек
Оценка аномальности	1 мин	3 мин	7 мин	12 мин

Табл. 1. Результаты оценки производительности.

Полученные результаты можно трактовать следующим образом. Время заполнения витрины данных и оценки аномальности фактов линейно зависит от количества фактов. Для полной загрузки и обработки данных, полученных с 10 тыс. компьютеров за месяц системе потребуется не более 80 часов (т.е. 12% времени работы подсистемы анализа). Учитывая тот факт, что анализ данных осуществляется в фоновом режиме и распределен во времени, можно утверждать, что одной подсистемы анализа так же достаточно для работы в сетях, насчитывающих более 10 тыс. компьютеров. Тиражируя подсистемы анализа можно обеспечить параллельную обработку данных крупных сетей, при этом анализ может осуществляться несколькими аналитиками.

Таким образом, предлагаемая в работе технология, основанная на построенной модели, разработанном системном решении и совокупности методов сбора, консолидации и анализа, позволяет создавать системы мониторинга, отвечающие всем поставленным требованиям.

Основные результаты работы

1. Предложена модель мониторинга, формализующая сбор и анализ данных, описывающих работу пользователей с информационными и вычислительными ресурсами корпоративной сети. Модель реализует концепцию раннего обнаружения внутренних вторжений и нецелевого использования ресурсов корпоративной сети за счет поиска отдельных аномалий в работе пользователей и изменений статистики работы.
2. Разработана технология построения специализированных систем мониторинга работы пользователей корпоративных сетей, реализующая предложенную модель. Предложенные в рамках технологии алгоритмические и программные решения позволяют учитывать особенности конкретных корпоративных сетей, обладают необходимой производительностью, обеспечивают надежность работы и защищенность данных.

Разработанная технология апробирована в виде экспериментальной системы мониторинга. Построенная система мониторинга официально внедрена в эксплуатацию в один из органов государственной власти Российской Федерации, насчитывающий более 1250 рабочих мест пользователей, и успешно эксплуатируется с 2009 года.

Публикации по теме диссертации

1. Трошин С. В. Консолидация и предобработка информации из журналов регистрации ВС для систем обнаружения вторжений // Сборник «Программные системы и инструменты», №6, М.:Изд-во «МАКС Пресс», 2005, сс. 68-80.
2. Машечкин И. В., Петровский М. И, Трошин С. В., Шестимеров А. А. Система сбора данных и анализа поведения пользователей вычислительной сети// Сборник тезисов XIV Международной конференции «ЛОМОНОСОВ», МГУ, 2007, сс. 79-80.
3. Igor V. Mashechkin, Mikhail I. Petrovskiy, Sergey V. Troshin, Andrew A. Shestimerov Data Gathering and User Behavior Analysis System // SYRCoSE 2007., 2007, volume 1, pp. 39-45.
4. Машечкин И. В., Петровский М. И., Трошин С. В., Шестимеров А. А. Система мониторинга и анализа поведения пользователей компьютерной системы // Сборник докладов Международной научной конференции «Системный анализ и информационные технологии» (САИТ-2007) , Обнинск, 2007, сс. 183-187.
5. Петровский М.И., Машечкин И.В., Трошин С.В. Исследование и разработка методов интеллектуального анализа данных для задач компьютерной безопасности// Сборник докладов Международной научной конференции «Математические методы распознавания образов» (ММРО-13), Москва, 2007, сс. 522-528.
6. Трошин С. В. Консолидация данных об активности пользователей вычислительных систем// Сборник тезисов XX Международной научной конференции «Математические методы в технике и Технологиях», Ростов-на-Дону, Изд. центр ДГТУ, 2007, сс. 254-256.
7. Машечкин И. В., Петровский М.И., Трошин С.В. Применение моделирования поведения пользователей в задачах компьютерной безопасности// Вестник молодых ученых «Ломоносов» выпуск IV, Москва 2007, сс. 87-94.
8. Машечкин И. В., Петровский М. И., Трошин С. В. Мониторинг и анализ поведения пользователей компьютерных систем // Сборник докладов VI Международной научной конференции по программированию УкрПРОГ’2008, Украина, г. Киев, 2008., сс. 22-29.
9. Трошин С. В. Мониторинг работы корпоративных пользователей //Вопросы современной науки и практики. Университет им. В. И. Вернадского №2(16)/2009, г. Тамбов, 2009, сс. 59-72.