(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2011년01월05일
(11) 등록번호 10-1005866
(24) 등록일자 2010년12월28일

(51) Int. Cl.
G06F 21/00 (2006.01) *G06F 15/00* (2006.01)
G06F 9/44 (2006.01)
(21) 출원번호 10-2008-0086906
(22) 출원일자 2008년09월03일
심사청구일자 2008년09월03일
(65) 공개번호 10-2010-0027836
(43) 공개일자 2010년03월11일
(56) 선행기술조사문헌
JP2005038116 A*
논문2(2007.05)*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
충남대학교산학협력단
대전광역시 유성구 궁동 220번지 충남대학교
이형우
경기도 용인시 기흥구 중동 성산마을서해그랑블 3101동 501호
한신대학교 산학협력단
경기 오산시 양산동 한신대학교내
(72) 발명자
이형우
경기도 용인시 기흥구 중동 성산마을서해그랑블 3101동 501호
(74) 대리인
특허법인태동

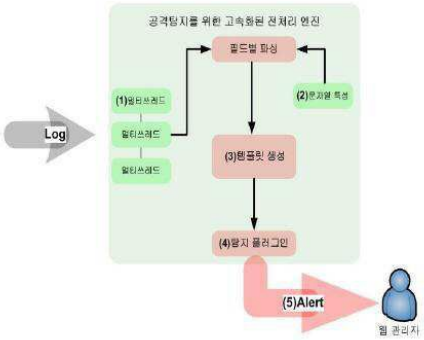
전체 청구항 수 : 총 14 항 심사관 : 권영학

(54) 룰기반 웹아이디에스 시스템용 웹로그 전처리방법 및 시스템

(57) 요 약

본 발명은 웹 로그 정보에 대한 효율적인 검색 기능을 제공하며 동시에 웹 서버에 의해 생성되는 대량의 로그 정보를 대상으로 한 룰 기반 공격 탐지의 효율성을 높이기 위해 전처리 과정을 수행하여 웹 IDS 시스템의 공격탐지 성능을 향상시킬 수 있도록 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법 및 시스템에 관한 것으로, 룰기반 웹아이디에스 시스템의 공격탐지를 위하여 일정한 포맷으로 구성된 웹로그를 전처리하는 방법에 있어서, 전체 로그 파일을 멀티쓰레드를 이용해 필드 단위로 분할하는 제1 단계와; 필드 단위로 분할된 로그에서 중복 문자열 인덱스 테이블을 구성하는 제2 단계와; 로그 분할 및 인덱싱 테이블을 이용하여 웹 로그 정보를 축약된 정보로 변환하여 템플릿(T_{WLX})을 생성하는 제3 단계를 포함하여 이루어진 것을 특징으로 한다.

대 표 도 - 도6



특허청구의 범위

청구항 1

룰기반 웹아이디에스 시스템의 공격탐지를 위하여 일정한 포맷으로 구성된 웹로그를 전처리하는 방법에 있어서,

전체 로그 파일을 멀티쓰레드를 이용해 필드 단위로 분할하는 제1 단계와;

필드 단위로 분할된 로그에서 중복 문자열 인덱스 테이블을 구성하는 제2 단계와;

로그 분할 및 인덱싱 테이블을 이용하여 웹 로그 정보를 축약된 정보로 변환하여 템플릿(T_{WLx})을 생성하는 제3 단계를 포함하여 이루어진 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법.

청구항 2

제 1 항에 있어서, 인덱싱 정보 기반 웹공격 탐지과정을 수행하고 결과를 제시하는 제4 단계를 더 포함하여 이루어진 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법.

청구항 3

제 1 항에 있어서, 상기 제1 단계는 아래의 수학식 1 및 수학식 2를 만족하는 로그 파일(D_{WLx})내 필드(fx)별로 분리하는 과정인 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법.

<수학식 1>

$$D_{WL_x} = \{L_1 | L_2 | \cdots | L_x\}$$

<수학식 2>

$$L_x = \{f_{date}, f_{time}, f_{SIP}, f_{DIP}, \ldots\}$$

.

청구항 4

제 1 항에 있어서, 상기 제2 단계는 필드 단위로 중복 문자를 제거하고 문자열을 키 값(kfx)으로 변환하는 제5 단계와;

검색 성능 향상을 위해 B-트리 기반 인덱스 테이블(I_{fx})을 구성하는 제6 단계를 포함하여 이루어지되,

상기 키 값(kfx) 및 B-트리 기반 인덱스 테이블은 아래의 수학식을 만족하는 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법.

<수학식 3>

$$I_{f_x} = [k_{f1} | k_{f2} | \cdots | k_{fx}]$$

.

청구항 5

제 1 항에 있어서, 상기 제3 단계의 템플릿은 아래 수학식을 만족하는 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법.

<수학식 4>

$$T_{WL_x} = \{I_{f_{date}} | I_{f_{time}} | I_{f_{sip}} | \cdots | L_{f_n}\}$$

청구항 6

삭제

청구항 7

제 1 항에 있어서, 제 1 단계에서 비교대상이 되는 로그파일의 필드는 원격지 IP 주소(%a), 헤더를 포함한 전송량(%b), 첫 번째 요청 라인(%r), 요청한 URL(%U)들인 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법.

청구항 8

제 1 항에 있어서, 상기 제2 단계는 문자열 특성을 고려한 중복 문자열 처리를 수행하되 원문로그 파일의 각 필드 단위로 구성된 테이블에 중복된 문자열을 인덱싱하는 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법.

청구항 9

룰기반 웹아이디에스 시스템의 공격탐지를 위하여 일정한 포맷으로 구성된 웹로그를 전처리하는 시스템에 있어서,

전체 로그 파일을 멀티쓰레드를 이용해 필드 단위로 분할하는 필드분할모듈과;

필드 단위로 분할된 로그에서 중복 문자열 인덱스 테이블을 구성하는 인덱스테이블구성모듈과;

로그 분할 및 인덱싱 테이블을 이용하여 웹 로그 정보를 축약된 정보로 변환하여 템플릿(T_{WLx})을 생성하는 템플릿생성모듈;을 포함하여 이루어진 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템.

청구항 10

제 9 항에 있어서, 인덱싱 정보 기반 웹공격 탐지과정을 수행하고 결과를 제시하는 탐지모듈을 더 포함하여 이루어진 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템.

청구항 11

제 9 항에 있어서, 상기 전체 로그 파일을 멀티쓰레드를 이용해 필드 단위로 분할하는 과정은 아래의 수학식 1 및 수학식 2를 만족하는 로그 파일(D_{WLx})내 필드(fx)별로 분리하는 과정인 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템.

<수학식 1>

$$D_{WL_x} = \{L_1 | L_2 | \cdots | L_x\}$$

<수학식 2>

$$L_x = \{f_{date}, f_{time}, f_{SIP}, f_{DIP}, \ldots\}$$

.

청구항 12

제 9 항에 있어서, 상기 필드 단위로 분할된 로그에서 중복 문자열 인덱스 테이블을 구성하는 단계는 필드 단위로 중복 문자를 제거하고 문자열을 키 값(kfx)으로 변환하고;

검색 성능 향상을 위해 B-트리 기반 인텍스 테이블(I_{fx})을 구성하는 과정을포함하여 이루어지되,

상기 키 값(kfx) 및 B-트리 기반 인덱스 테이블은 아래의 수학식을 만족하는 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템.

<수학식 3>

$$I_{f_x} = [k_{f1} | k_{f2} | \cdots | k_{fk}]$$

.

청구항 13

제 9 항에 있어서, 상기 로그 분할 및 인덱싱 테이블을 이용하여 웹 로그 정보를 축약된 정보로 변환하여 템플릿(T_{WLX})을 생성하는 단계는 템플릿이 아래 수학식을 만족하는 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템.

<수학식 4>

$$T_{WL_x} = \{I_{f_{date}} | I_{f_{time}} | I_{f_{ip}} | \cdots | L_{f_n}\}$$

청구항 14

삭제

청구항 15

제 9 항에 있어서, 상기 전체 로그 파일을 멀티쓰레드를 이용해 필드 단위로 분할하는 단계에서 비교대상이 되는 로그파일의 필드는 원격지 IP 주소(%a), 헤더를 포함한 전송량(%b), 첫 번째 요청 라인(%r), 요청한 URL(%U)들인 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템.

청구항 16

제 9 항에 있어서, 상기 필드 단위로 분할된 로그에서 중복 문자열 인덱스 테이블을 구성하는 단계는 문자열 특성을 고려한 중복 문자열 처리를 수행하되 원문로그 파일의 각 필드 단위로 구성된 테이블에 중복된 문자열을 인덱싱하는 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템.

명 세 서

발명의 상세한 설명

기 술 분 야

[0001] 본 발명은 룰기반 웹아이디에스 시스템용 웹로그 전처리방법 및 시스템에 관한 것으로, 더욱 상세하게는, 웹 로그 정보에 대한 효율적인 검색 기능을 제공하며 동시에 웹 서버에 의해 생성되는 대량의 로그 정보를 대상으로 한 룰 기반 공격 탐지의 효율성을 높이기 위해 전처리 과정을 수행하여 웹 IDS 시스템의 공격탐지 성능을 향상시킬 수 있도록 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 국내의 인터넷 이용률은 꾸준히 증가 추세에 있으며, 현재 국내 인터넷 사용자 수는 약 34,570(천명)을 넘고 있다. 그리고 인터넷 사용자를 대상으로 한 조사에서 전체 사용자의 73.2%가 하루 1회 이상 인터넷을 사용하고 있는 추세이다.

[0003] 이처럼 인터넷 사용자의 증가에 따라 국내 주요 포탈 웹 사이트의 하루 웹 로그 양은 50G 내외로 대용량의 로그가 발생하고 있다. 그러나 웹의 양적인 증가와 더불어 웹 공격의 시도 및 성공도 함께 증가하고 있다. 이처럼 웹 서비스가 해커들의 공격 대상이 된 이유는 웹 서비스와 웹 어플리케이션의 빠른 증가 추세와 함께 비즈니스 및 많은 사업군이 웹 기반 서비스 방식으로 변화되어 웹 시스템에 대한 의존도가 높아졌기 때문이다. 또한 웹 서비스의 특성상 특정포트(예를 들면, 80과 443번 포트)를 통해 외부로부터 유입되는 접속을 허용할 수밖에 없기 때문에 방화벽 시스템에 다른 포트가 차단하였다 하더라고 공격자는 손쉽게 웹 서버에 대한 공격을 수행할 수 있다.

[0004] 이러한 웹 서비스의 취약점을 보완하기 위해 현재 웹 서비스를 대상으로한 공격 탐지 시스템(Web IDS : Web Intrusion Detection System)이 제시되었다. 기존의 IDS 시스템에서는 공격을 탐지하기 위해 IP 패킷을 대상으로 룰 데이터를 이용해 공격 여부를 탐지한다. 그러나 웹 서버의 공격 탐지를 위해서는 웹 서버에서 생성되는 웹 로그(Web Log) 데이터에 대한 분석을 통해 외부로부터의 불법적인 접속을 탐지하거나 이상 탐지(Anomaly Detection) 기능을 제공해야 한다. 최근 데이터 마이닝 기술을 적용하여 웹 공격에 대한 이상탐지 및 대응 시스템 구축에 활용하는 방안에 대한 연구도 진행되고 있다. 데이터 마이닝 기법에서의 성능 향상을 위해서는 대량

의 웹 로그에 대한 효율적 전처리 기법이 제시되어야 한다.

[0005] 기존의 웹 IDS 시스템은 웹 로그를 기반으로 외부로부터의 공격이나 웹 시스템 내부의 부적절한 쿼리 전송 및 이상 접속 정보를 탐지하기 위해 공격 탐지 룰(Web Attack Rule) 정보를 사용한다. 하지만 기존의 웹 IDS 시스템은 대량으로 생성되는 웹 로그에 대한 별도의 전처리 과정 없이 웹 공격 탐지 룰과 비교하는 방식이므로 실시간으로 수행되는 웹 공격에 효율적으로 대처하지 못하고 있다.

[0006] 한편, 웹 사용자의 급증과 동시에 사용자의 개인정보 유출 및 기업 홈페이지의 변조, 금융사고 등과 같은 웹 해킹 사고가 급증하고 있다. 이러한 사건들은 대부분 시스템을 해킹하는 것이 아니라 누구에게나 개방되어 있는 홈페이지를 통해서 시스템에 침투하는 것이다.

[0007] 웹을 이용한 공격 기법은 꾸준히 고도화되고 있으며 공격 빈도도 급격히 증가하고 있어 이에 대한 대응 기법이 제시되어야 한다. 현재까지 공개된 웹 관련 공격 형태는 알려진 것만 무려 4,000개 이상이고 알려지지 않은 공격을 포함한다면 그 수는 헤아릴 수 없는 상황이다. 특히 최근에는 세션 하이재킹, 패킷 스니핑 및 스캔 기술을 중심으로 한 네트워크 공격 기술에서 전반적으로 웹 해킹 및 공격 기술이 급증하고 있는 추세이다. 웹 서비스에 대한 공격 기법을 살펴보면 다음과 같다.

[0008] 웹 서비스는 버퍼 오버플로우(buffer overflow), 세션 하이제킹 및 SQL 주입/파라메터 주입 공격 등이 가능하여 최근 웹 서버에 대한 DoS 공격으로 발전하고 있다. 대표적인 웹 공격 방식인 크로스 사이트 스크립트 취약점 공격, SQL 주입(SQL Injection) 공격 취약점 공격 방식에 대해 살펴보면 다음과 같다.

[0009] 크로스 사이트 스크립트(XSS) 공격

[0010] XSS 공격 기법은 JavaScript, VBScript, Flash, ActiveX, XML/XSL, DHTML 등과 같이 클라이언트 측에서 실행되는 언어로 작성된 코드를 사용자 입력으로 주게 되면 이 코드가 그대로 클라이언트 측 브라우저에서 수행되는 특성을 이용해 악성 스크립트 코드를 웹 페이지, 웹 게시판, 웹 메일에 포함시켜 사용자에게 전송하게 된다. 예를 들면 웹 사용자가 취약한 웹서버에 접속 중일 때 공격자는 악성 스크립트를 업로드한 후 웹 사용자에게 악성 스크립트가 있는 링크를 클릭하도록 유도한다. 웹 사용자가 해당 링크를 클릭하게 되면 자신의 쿠키 등의 정보가 공격자에게 전송된다. 공격자는 수집된 정보를 이용해 피해자의 권한을 획득해 피해자의 권한으로 웹서버를 사용할 수 있게 된다. 이때 사용되는 스크립트들은 해커들에 의해 제작되고 배포되어 일반인들도 쉽게 사용할 수 있다.

[0011] SQL 주입 공격

[0012] 현재 대부분의 웹 사이트들은 사용자로부터 입력받은 값을 이용해 DB 접근을 위한 SQL 쿼리(query)를 만들고 있다. 사용자 로그인 과정을 예로 들면, 사용자가 유효한 계정과 패스워드를 입력했는지 확인하기 위해 사용자 계정과 패스워드에 관한 SQL 쿼리문을 만든다. 이때 SQL 주입 공격 기법을 통해서 정상적인 SQL 쿼리를 변조할 수 있도록 조작된 사용자 이름과 패스워드를 보내 정상적인 동작을 방해할 수 있다. 이러한 비정상적인 SQL 쿼리를 이용해 다음과 같은 공격이 가능하다.

[0013] - 사용자 인증을 비정상적으로 통과

[0014] - 데이터베이스에 저장된 데이터를 임의로 열람

[0015] - 데이터베이스의 시스템 명령을 이용하여 시스템 조작

[0016] DoS(Denial of Service) 공격

[0017] DoS 공격은 특정 시스템에 대한 불법적인 권한을 얻는 적극적인 방법이 아니라 네트워크와 시스템의 자원을 공격 대상으로 하는 공격방법이다. 웹에서의 DoS 공격에 대해서는 패킷의 발신지 IP를 중심으로 특정 서버에 대한 자원의 요청 및 오류 메시지의 발생 빈도를 통해 공격을 탐지할 수 있다. 그러나 현재는 IP 스푸핑과 같은 공격으로 공격의 근원지 역추적이 힘든 실정이다. 스푸핑 방식을 통한 DoS 공격은 사용자 ID 정보를 중심으로 전처리 모듈에서 IP 스푸핑 공격 여부를 확인 할 수 있으며, 웹 서비스의 취약점을 보완하기 위해 룰 기반의 웹 IDS

시스템이 제시되었다. 하지만 기존의 룰 기반의 웹 IDS 시스템은 다음과 같은 취약점을 보이고 있어 이에 대한 대응 기술이 제시되어야 한다.

[0018] 기존의 룰 기반의 웹 IDS 및 취약점

[0019] HTTP 프로토콜은 가장 범용적으로 사용되는 프로토콜중의 하나로 많은 지식을 필요로 하지 않으며 URL 상의 간단한 조작 및 유추 등으로도 웹 해킹이 가능하다는 특징이 있다. 따라서, 기존의 일반적 형태의 IDS 시스템만으로는 웹 서비스에 특화된 공격에 능동적으로 대응할 수 없게 되었기에 웹 IDS 시스템이 필요하게 되었다.

[0020] 현재 룰 기반의 웹 IDS는 웹 서버 장치 각각의 공격 탐지를 위해 설계된 HIDS (Host-IDS)이다. 이때 사용되는 룰은 도 1과 같은 형식이다. 룰 기반의 웹 IDS는 도 2와 같이 시스템 내부의 로그 파일을 분석해 정의된 룰과의 비교를 통해 공격을 탐지하는 기술이다.

[0021] 기존의 전처리 기술은 다음 단계를 통해 공격을 탐지하여 관리자에게 탐지 결과를 전달했다.

[0022] 1단계: 로그 수집

[0023] 2단계: 수집된 로그와 룰과의 비교

[0024] 3단계: 룰 적용 결과를 관리자에게 전송

[0025] 그러나 기존의 웹 IDS는 로그 파일의 전처리 과정이 없이 순차 검색을 통해 룰과의 비교를 수행하고 있다. 따라서 최악의 경우 웹 로그 생성 후 일정 시간이 경과된 이후에서야 탐색 결과를 제시하게 된다는 단점이 있다. 기존의 시스템에서는 가공되지 않은 원본 형태의 웹 로그를 대상으로 룰 기반 공격 탐지를 수행하기 때문에 만족할만한 성능을 보이지 못하고 있다.

[0026] 기존 웹 로그 전처리 기술의 취약점

[0027] 기존 웹 로그 전처리 기술

[0028] 웹 로그는 일정한 형태의 포맷으로 구성되어 진다. 그 종류로는 일반적인 형태의 CLF(Common Log Format) 형태와 확장된 로그 파일 형식인 ELF(Extended Log Format)로 크게 구분 지을 수 있다. 위 포맷은 일반적으로 도 3과 같은 형태로 구성되며, 각 필드별로 저장된 정보를 이용하여 웹 IDS 시스템 기반 공격탐지 및 이상탐지 기능에 적용 가능한 형태로 변형되어야 한다.

[0029] 기존 웹 로그 전처리 기술에서는 다음 과정을 수행해 사용자 성향 및 웹 분석을 위해 전처리 과정을 수행하였다.

[0030] - 1단계 : 로그 수집 - 현재 웹 로그의 포맷은 일반적으로 가장 많이 사용되는 CLF 포멧, 확장된 로그 파일 형식의 ELF 포맷 형식으로 크게 나눌 수 있다. ELF 형식은 다시 MicroSoft사의 웹서버에서 사용하는 MS-IIS 포멧과 NCSA 계열의 웹서버에서 사용하는 NCSA 포맷으로 나눌 수 있다.

[0031] - 2단계 : Cleaning Log- 분석에 필요하지 않은 아이템을 정제하며 일반적으로 gif, jpeg, jpg, map 등을 로그 파일에서 삭제함. 데이터 용량이 일반적으로1/10에서 1/40 정도로 축소되는 효과를 가져온다. 그러나 현재의 웹 공격 성향을 분석한 결과 이미지 파일로 가장된 백도어 프로그램을 유포시키는 등 기존의 우회 공격이 존재하고 있다.

[0032] - 3단계 : User Identification- 로그 파일에 기록된 사용자 정보를 확인하여 이동경로를 정제하는 과정이다. 일반적으로 IP Address와 Browser의 환경 정보를 이용하여 전처리한다.

[0033] - 4단계 : Session Identification- 사용자의 세션 초과 유무를 점검하여 정제하는 과정이다. 임계치 값을 설정하여 사용자 세션 분류 과정을 수행한다(도 4참조).

[0034] - 5단계 : Path Completion- 로그에 기록되지 않은 이동 경로 정보를 연결하는 과정이다. 일반적으로 back 또는 forward 버튼을 눌러 이동한 경우 경로 연결 과정을 수행한다(도 5 참조).

[0035] - 6단계 : Formatting- 웹 로그 분석에 적합한 정보의 형태로 포맷을 전환한다. 포맷 형태는 사용되는 데이터

분석 기법에 따라 결정한다.

[0036] 초고속 인터넷의 발달로 웹 사용자들이 급격히 증가하였고, 많은 웹 서비스들은 사용자들에게 공개되고 있다. 이런 공개된 웹 서비스에 대해 최근 해커들의 공격 역시 급격히 증가하고 있는 추세이다.

[0037] 그러나 기존의 웹 로그 전처리 및 분석 기법은 웹 서비스 사용자 성향 및 이용 경로 분석을 위한 고전적 전처리 방식을 그대로 사용하고 있어 다음과 같은 문제점을 가지고 있다.

[0038] - 로그내 각 필드 정보의 문자열 특성을 배제하고 단순히 문자열 대상 순차 탐색 방식만을 사용하여 전처리 및 공격 탐지 수행

[0039] - 대용량의 웹 로그 정보에 대한 고속처리 기능이 미비

[0040] - 웹 공격 탐지를 위한 효율적인 자료구조미비 및 검색 성능 향상을 위한 템플릿 생성 모듈 미비.

발명의 내용

해결 하고자하는 과제

[0041] 본 발명의 목적은 상기한 바와 같은 종래기술의 문제점을 해결하기 위해 제안된 것으로 웹 로그 정보에 대한 효율적인 검색 기능을 제공하며 동시에 웹 서버에 의해 생성되는 대량의 로그 정보를 대상으로 한 룰 기반 공격 탐지의 효율성을 높이기 위해 전처리 과정을 수행하여 웹 IDS 시스템의 공격탐지 성능을 향상시킬 수 있도록 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법 및 시스템을 제공하고자 하는 것이다.

과제 해결수단

[0042] 상기한 바와 같은 목적을 달성하기 위한 본 발명의 바람직한 실시예에 의하면, 룰기반 웹아이디에스 시스템의 공격탐지를 위하여 일정한 포맷으로 구성된 웹로그를 전처리하는 방법에 있어서,

[0043] 전체 로그 파일을 멀티쓰레드를 이용해 필드 단위로 분할하는 제1 단계와;

[0044] 필드 단위로 분할된 로그에서 중복 문자열 인덱스 테이블을 구성하는 제2 단계와;

[0045] 로그 분할 및 인덱싱 테이블을 이용하여 웹 로그 정보를 축약된 정보로 변환하여 템플릿(T_{WLx})을 생성하는 제3 단계를 포함하여 이루어진 것을 특징으로 하는 룰기반 웹아이디에스 시스템용 웹로그 전처리방법이 제공된다.

[0046] 본 발명에 따른 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템에 따르면, 룰기반 웹아이디에스 시스템의 공격탐지를 위하여 일정한 포맷으로 구성된 웹로그를 전처리하는 시스템에 있어서, 전체 로그 파일을 멀티쓰레드를 이용해 필드 단위로 분할하는 필드분할모듈과; 필드 단위로 분할된 로그에서 중복 문자열 인덱스 테이블을 구성하는 인덱스테이블구성모듈과; 로그 분할 및 인덱싱 테이블을 이용하여 웹 로그 정보를 축약된 정보로 변환하여 템플릿(T_{WLx})을 생성하는 템플릿생성모듈;을 포함하여 이루어진 것을 특징으로 한다.

효 과

[0047] 이상 설명한 바와 같이, 본 발명에 따른 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템 및 방법에 의하면, 로그 정보를 각 필드 별로 분할한 후 B-Tree 기반의 룰 기반 탐색 과정을 수행하도록 하여 기존의 방법보다 대량의 로그 정보를 고속으로 처리하게 되며 룰 기반 탐지 모듈과 연계시킬 수 있기 때문에 웹 서비스 공격을 보다 효과적으로 탐지 및 대처할 수 있는 효과가 있다.

[0048] 또한 본 발명에 따르면, 웹 로그에 대한 전처리 과정을 수행할 경우 앞에서 제시한 웹 서비스의 취약점들이 이용한 스크립트 업로드, 쿼리 전송 등 웹 서버에서의 공격 행위들에 대해 효율적으로 탐지할 수 있는 효과가 있다. 이를 통해 대용량 웹 로그 정보에 대한 효율적인 검색 기능을 제공하며 동시에 웹 로그에 대한 룰기반 IDS 시스템에 적용 가능하여 공격 탐지 성능을 향상시킬 수 있는 효과가 있다.

발명의 실시를 위한 구체적인 내용

[0049] 이하 본 발명에 따른 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템 및 방법을 첨부도면을 참조로 상세히

설명한다.

[0050] 본 발명에 따른 룰기반 웹아이디에스 시스템용 웹로그 전처리시스템 및 방법방법의 모듈별 전체적인 흐름도는 도 6과 같다.

[0051] *1 단계(멀티쓰레드): 전체 로그 파일을 멀티쓰레드를 이용해 필드 단위로 분할함

[0052] - 로그 파일(D_{WLx})내 필드(fx)별로 분리(필드분리모듈)

[0053] <수학식 1>

[0054] $$D_{WL_x} = \{L_1 | L_2 | \cdots | L_x\}$$

[0055] <수학식 2>

[0056] $$L_x = \{f_{date}, f_{time}, f_{SIP}, f_{DIP}, \ldots\}$$

[0057] 2단계(문자열 특성): 필드 단위로 분할된 로그에서 중복 문자열 인덱스 테이블을 구성(인덱스테이블구성모듈)

[0058] - 필드 단위로 중복 문자를 제거하고 문자열을 키 값(kfx)으로 변환

[0059] - 검색 성능 향상을 위해 B-트리 기반 인덱스 테이블(I_{fx})을 구성함

[0060] <수학식 3>

[0061] $$I_{f_x} = [k_{f1} | k_{f2} | \cdots | k_{fx}]$$

[0062] - 3단계(템플릿 생성): 로그 분할 및 인덱싱 테이블을 이용하여 웹 로그 정보를 축약된 정보로 변환하여 템플릿(T_{WLx})을 생성(템플릿생성모듈)

[0063] <수학식 4>

[0064] $$T_{WL_x} = \{I_{f_{src}} | I_{f_{time}} | I_{f_{sip}} | \cdots | L_{f_n}\}$$

[0065] - 4단계(탐지 플러그인): 인덱싱 정보 기반 웹 공격 탐지 과정을 수행 및 결과 제시(탐지모듈)

[0066] 웹 IDS 시스템에 접목시키기 위해서 웹 로그(D_{WLx}) 정보내 문자열 특성을 고려한 중복 문자열 처리 방식은 원문 웹 로그 파일의 각 필드(fx) 단위별로 구성된 테이블(I_{fx})에 중복된 문자열을 인덱싱(indexing)하는 방식이다.

[0067] 또한 본 발명에서는 도 7과 같이 웹 로그내 중복 문자열 인덱스 처리 과정을 수행하므로 원문 웹 로그보다 물리적인 효율성을 제공한다. 예를 들어 웹 로그의 특정 필드에 특정 문자열을 검색할 때 문자열의 검색보다는 인덱스된 키 값(kfx)을 통해 검색하므로 효율적이다. 만일 키워드 검색 과정을 수행하고자 할 경우, 전체 로그 파일의 각 필드 정보가 이미 인덱스 정보로 저장되어 있기 때문에 SQL 주입 공격 등과 같은 웹 공격 시도에 대해 빠른 검색 기능을 제공하며, 인덱스 정보로부터 본래의 웹 로그 정보로 변환 생성이 가능하다.

[0068] 즉, 본 발명에서 제시한 방법은 원문 로그의 한 라인이 가지고 있는 각 필드의 인덱스를 통해 B-트리를 거치지 않고 해당 필드의 문자열들을 가져와 바로 하나의 로그 열로 통합하여 다시 원문 로그로 복구시킬 수 있다.

[0069] 각 단계별로 수행되는 과정에 대해 설명하면 다음과 같다.

[0070] 구체적으로 도 8과 같이 멀티쓰레드 기법을 사용하여 원문 로그를 분할 통치하는 알고리즘 기법을 사용한다. 대량의 웹 로그 정보에 대해 멀티쓰레드 방식으로 분할하여 전처리 과정을 수행한다. 각 쓰레드에 할당되는 정보는 대용량 웹 로그내 동일 필드별 정보를 대상으로 전처리 과정을 수행하도록 하였다.

[0071] 제안하는 전처리 기법은 로그 파일을 분할하여 룰 비교를 효율적으로 할 수 있게 중복 문자열을 처리한다. 이때

생성되는 템플릿은 도 9와 같은 구조로 생성된다.

[0072] 본 발명서 비교 대상이 될 로그 파일의 필드는 원격지 IP 주소(%a), 해더를 포함한 전송량(%b), 첫 번째 요청 라인(%r), 요청한 URL(%U)들이다. 각 필드별로 나누어진 로그 데이터는 본 발명에서 제안하는 멀티쓰레드 기반 전처리 모듈에 의해 처리된다.

[0073] 문자열 특성을 고려한 중복 문자열 처리는 원문 로그 파일의 각 필드 단위로 구성된 테이블에 중복된 문자열을 인덱싱하는 기법이다.

[0074] 각 필드별로 분할된 로그 파일들은 중복 문자열 처리 과정을 수행한 결과 약 50%의 성능향상을 가져올 수 있으며, 전체 로그 파일의 각 필드 정보들은 인덱스 정보들로 저장되어 완전한 문자열로의 변환/생성이 가능하다.

[0075] 기존의 룰 기반의 웹 IDS는 access.log 파일의 순차적인 문자열 탐색을 통해 웹 공격을 탐지했다. 그러나 웹 공격의 입력 값 부재 공격은 클라이언트가 서버에게 요청하는 쿼리문에 의해 공격이 이루어진다. 그러므로 본 발명에서 제안하는 방법은 룰 비교 시 불필요한 탐색 과정을 분할 통치법을 이용해 각각의 필드별로 분할하고 룰과의 비교를 이진 탐색 알고리즘을 이용해 기존 알고리즘을 개선한다.

[0076] 크기가 n인 선형 리스트에서 원소들의 키 값이 주어진 키 값과 같을 확률이 1/n 로 모두 같다고 할 때 임의의 위치 k에서 탐색키를 찾는데 k번 비교 연산이 필요하며, 평균 비교 횟수는 수학식 5와 같다.

[0077] <수학식 5>

[0078] $$\sum_{k=1}^{n} k\frac{1}{n} = \frac{n+1}{2}$$

[0079] 제안 기법인 이진 탐색 법은 한번 탐색시마다 탐색 원소의 개수가 반으로 줄어든다. 수학식 6은 n이 거듭제곱수라 할 때 최악의 실행시간 T(n)이다.

[0080] <수학식 6>

[0081] $$\begin{aligned} T(n) &= T(\tfrac{n}{2}) + \Theta(1) \\ &= T\tfrac{n}{2^2} + \Theta(1) + \Theta(1) \\ &= \cdots \\ &= T(\tfrac{n}{2^{\lg n}} + \Theta(1) + \Theta(1) + \cdots + \Theta(1) \\ &= \Theta(\lg n) \end{aligned}$$

[0082] n이 거듭제곱 수가 아닌 경우의 이진 탐색법의 시간 복잡도는 o(logn)이다.

[0083] 본 발명에서 제안한 룰 비교시 탐색 기법은 실제 비교 대상의 문자열들이 정렬되어 있어 실제 룰과의 비교는 이진 트리의 탐색에서 이루어진다. SQL 주입 공격 및 파라메터 주입 공격은 클라이언트가 요청한 url의 비교를 통해 공격 탐지가 가능하다. 그러므로 기존의 룰 중 url의 문자열을 비교하여 공격을 탐지한다.

도면의 간단한 설명

[0084] 도 1은 종래의 룰기반 웹아이디에스 시스템의 필드형식을 나타낸 것이다.

[0085] 도 2는 종래의 룰기반 웹아이디에스 시스템에서의 공격탐지과정을 나타낸 개념도이다.

[0086] 도 3은 종래의 웹로그 구성 필드를 나타낸 것이다.

[0087] 도 4는 종래의 웹로그 전처리 과정중 세션 인식(IDentification)과정을 나타낸 개념도이다.

[0088] 도 5는 종래의 웹로그 전처리 과정중 경로 완성(Path Completion)과정을 나타낸 개념도이다.

[0089] 도 6은 본 발명에 따른 웹로그 전처리모듈별 흐름도이다.

[0090] 도 7은 본 발명에 따른 웹로그 전처리 과정을 나타낸 흐름도이다.

[0091] 도 8은 본 발명에 따른 웹로그 전처리방법에서 사용되는 멀티쓰레드 기반고속처리 구조를 나타낸 도면이다.

[0092] 도 9는 본 발명에 따른 웹로그 전처리방법에서 생성된 탬플릿의 예를 나타낸 도면이다.

[0093] 도 10은 본 발명에 따른 웹로그 전처리방법에서 원문 로그의 인덱싱과정을 나타낸 도면이다.

[0094] 도 11은 본 발명에 따른 웹로그 전처리방법에서 중복 문자열 단위로 인덱싱된 필드를 나타낸 도면이다.

[0095] 도 12는 본 발명에 따른 웹로그 전처리방법에서 웹로그내 공격 탐색과정을 나타낸 도면이다.

도면

도면1

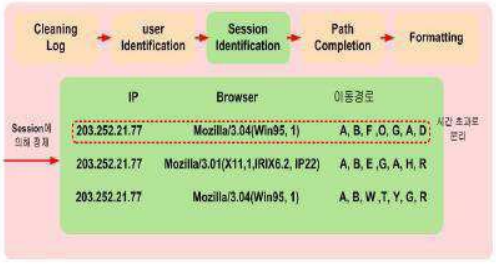
형식	예시
rule id	31103
if_sid	31100
url	'|select%20|select+|insert%20|%20from%20|%20where%20|union%20|
descripton	SQL injection
group	attack, sql_injection

도면2

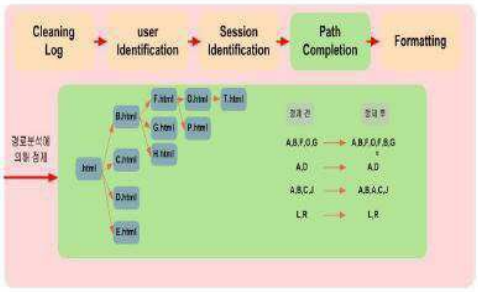
도면3

로그 구성 필드	설명
%a	원격의 IP 주소
%b	헤더를 포함한 전송량(byte)
%{var}e	환경변수"var"
%f	파일 이름
%h	원격의 호스트
%{hdr}i	서버에 들어오는(요청) 헤더 값 "hdr"
%{hdr}o	응답 헤더 값 "hdr"
%l	원격의 로그인 ID(지원한다면)
%{label}n	다른 모듈에서 "label"구성
%p	서버의 Canonical 포트 번호
%P	자식 Process ID(PID)
%r	첫 번째 요청 라인
%t	시간 포맷(CLF 포맷)
%{format}t	"format"으로 구성된 시간 포맷
%T	서버에 요청하는 시간(초)
%u	원격지의 유저 이름(인증시)
%U	요청한 URL
%v	클라이언트 요청에 따른 Canonical 서버 네임
%V	Use CanonicalName 설정에 따른 서버 네임

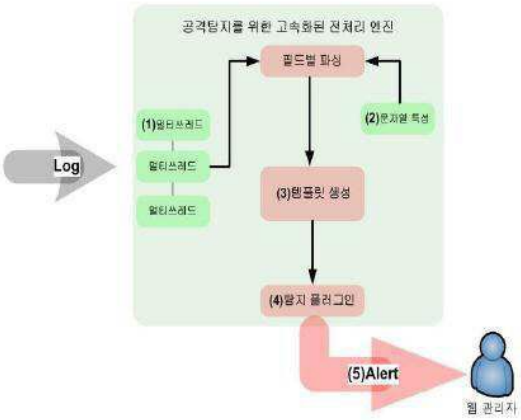
도면4



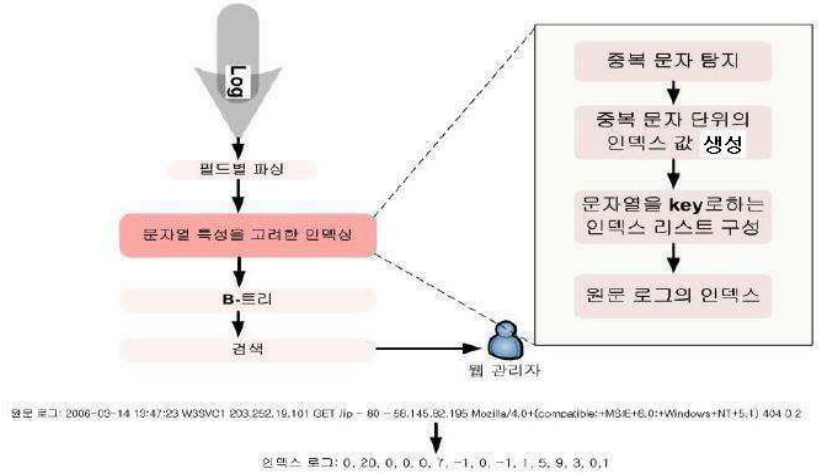
도면5


Cleaning Log
user Identification
Session Identification
Path Completion
Formatting

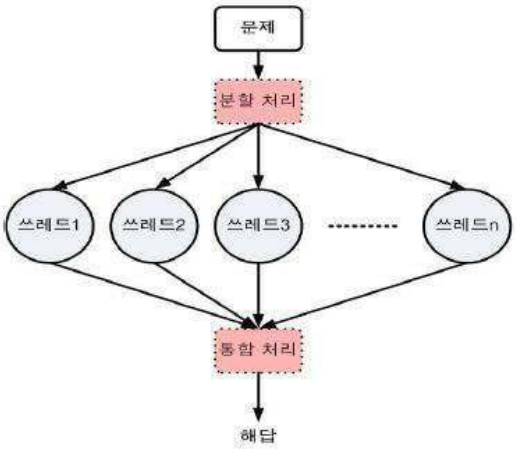
도면6


공격탐지를 위한 고속화된 전처리 엔진
필드별 파싱
(1)멀티쓰레드
멀티쓰레드
멀티쓰레드
(2)문자열 특성
(3)템플릿 생성
(4)탐지 플러그인
Log
(5)Alert
웹 관리자

도면7


Log
필드별 파싱
문자열 특성을 고려한 인덱싱
B-트리
검색
웹 관리자
중복 문자 탐지
중복 문자 단위의 인덱스 값 생성
문자열을 key로하는 인덱스 리스트 구성
원문 로그의 인덱스
원문 로그: 2006-03-14 13:47:23 W3SVC1 203.252.19.101 GET /ip - 80 - 58.145.82.195 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) 404 0 2
인덱스 로그: 0, 20, 0, 0, 0, 7, -1, 0, -1, 1, 5, 9, 3, 0,1

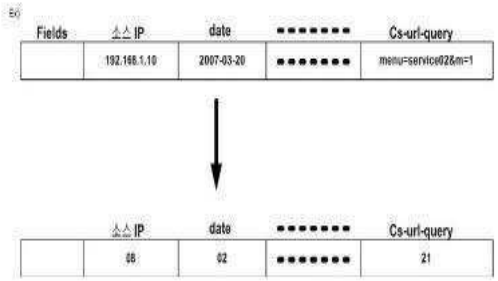
도면8


문제
분할 처리
쓰레드1
쓰레드2
쓰레드3
쓰레드n
통합 처리
해답

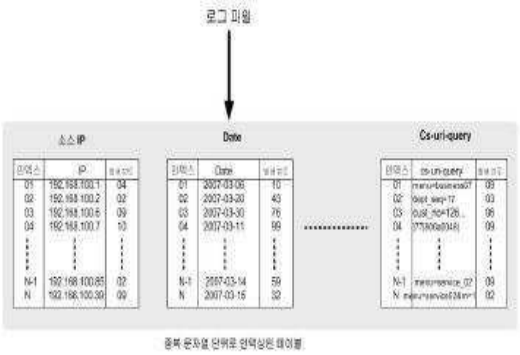
도면9

인덱스	cs-uri-query	발생 빈도
1	menu=business07	9
2	dept_seq=17	3
3	cust_no=126	6
4	|77|800a0046|	9
·	·	·
n-1	menu=service_02	9
n	menu=service02&m=1	2

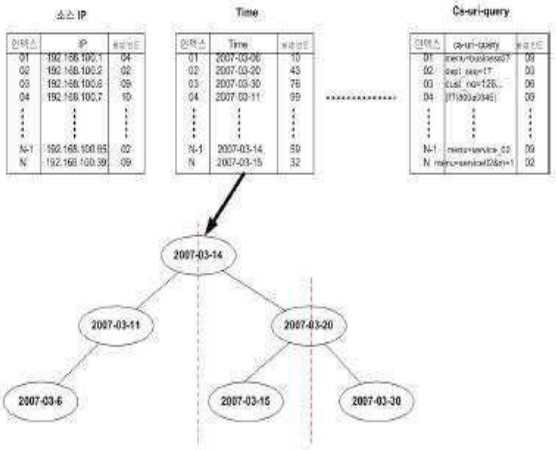
도면10



도면11



도면12


소스 IP
Time
Cs-uri-query
2007-03-14
2007-03-11
2007-03-20
2007-03-6
2007-03-15
2007-03-30