웹 로그를 이용한 실시간 로봇 탐지 알고리즘

오명진[a], 권신일[a0], 유준범[b], 차성덕[a]
[a] 고려대학교 컴퓨터학과, [b] 건국대학교 컴퓨터공학부
{bluehven, darkkal, scha}@korea.ac.kr, jbyoo@konkuk.ac.kr,

Real-time web robot detection algorithm using a web log

Myungjin Oh[a], Shinil Kwon[a0],Junbeom Yoo[b], Sungdeok Cha[a]
[a] Dept. of CS and Engineering. Korea Univ.
[b] Div. of CS and Engineering. Konkuk Univ.

1. 서 론

웹은 다양한 정보가 웹 서비스에 집중되면서 정보 전달의 메카로서 자리 매김 했을 뿐만 아니라, 각종 커뮤니티와 메신저 등의 발달로 이용 연령층이 다양해지면서 대중화에도 성공했다. 또한 쉬운 접근성과 각종 상업적인 서비스의 발전에 힘입어 웹을 통한 금전적 거래가 폭발적으로 증가하면서 웹은 현대 사회의 필수공간으로 각광 받고 있다.

웹에서의 로봇은 웹 상에서 동작하는 프로그램으로 순수한 연구 목적으로 만들어진 것에부터 상업적인 용도로 쓰이는 것까지 그 종류가 매우 다양하다. 이러한 로봇은 다음과 같은 이유로 실시간 탐지가 필요하다. 첫째, 5%의 로봇 세션이 전체 HTTP 요청의 85%를 차지한다는 연구[1]결과를 볼 때 웹 트래픽의 상당 부분을 로봇이 차지한다는 것을 알 수 있다. 웹 서버 관리자가 원하지 않은 로봇의 활동은 네트워크 성능을 떨어뜨려 사용자의 이용을 불편하게 한다. 둘째, 순수한 연구 목적으로 활용되거나 사용자의 단순 반복 작업을 대신해 주는 유용한 로봇도 있지만 과도한 트래픽 유발이나 해킹 등의 목적으로 악용되는 로봇도 존재한다. 웹의 대중성과 상업성을 고려해볼 때 로봇의 악의적인 공격은 심각한 개인 정보 침해나 막대한 재산 손실을 초래할 수 있다. 웹의 접근성과 전파력에 비추어 볼 때, 사고 발생시 수습이 상당히 어렵다. 따라서 웹에서 활동하고 있는 로봇의 실시간 탐지는 매우 중요하다

본 연구에서는 획득된 데이터의 신뢰도를 보장 하기 위해 Microsoft의 웹 로그를 사용하였다. 이는 기존의 유사한 연구와 차별되는 부분으로 기존 연구에서 사용된 데이터는 일반적으로 인터넷 쇼핑 사이트나 일반 대학의 로그 파일에서 획득한 것이다. 본 연구에서 사용된 웹 로그는 Microsoft와의 연구 협약을 통해 제공 받은 것으로, 전세계를 대상으로 서비스하는 MS 웹 페이지의 하루치 분량이다. 이는 10억 트랜젝션 이상의 Request 정보를 가지고 있으며 전체 로그의 용량이 250GB에 달한다. MS 웹 로그에는 엄청난 수의 일반 사용자 정보뿐만 아니라 다양한 종류의 로봇의 정보 또한 가지고 있기 때문에 로봇에 대한 연구를 검증 하는데 있어 충분한 신뢰도를 보장한다.

2. 본 론

웹의 중요성이 증가하면서 로봇 탐지 분야가 매우 주목 받고 있음에도 불구하고, 이 분야에 대한 연구가 많이 이루어지고 있지 않다. 그러나 몇 개의 주요한 연구가 공통적으로 제시하는 것은 일반 사용자와 로봇을 구분 짓는 몇 개의 특성이 존재한다는 것이다. 마찬가지로 본 연구에서도 일반 사용자와 로봇을 구별하는 결정적인 요소가 있다는 가정으로 실험하였다. 본 연구에서 중요하게 고려 된 특성은 대용량 웹 서버를 효과적으로 분류하기 위해 선행 된 연구[2,3]를 통해 선정하였고, 이는 IIS standard format을 기반으로 획득된 Client-Server Bytes, HEAD Request, Referrer ratio, Resource Type ratio, Error of Request, Robots.txt, 그리고 popularity of pages 이다.

본 논문에서는 실시간 로봇 탐지를 위한 Composite Attribute Vector(이하 CAV) 알고리즘을 제안한다. CAV 알고리즘은 로봇과 일반 사용자를 구별 할 수 있는 특징들을 각각의 벡터로 표현하고 표현된 각 특성 벡터들을 평준화하여 합성함으로써 로봇과 일반 사용자를 구별한다. 특별히 본 논문에서는 이러한 복합특성벡터를 CAV라 표현한다.

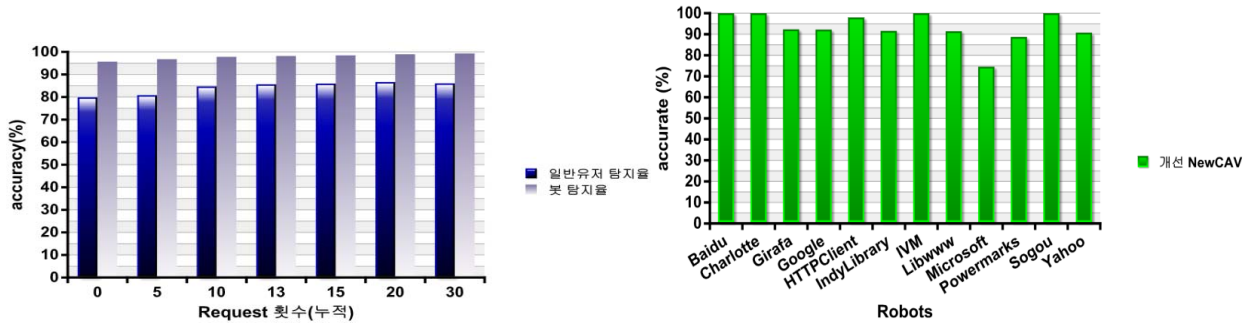

그림. CAV 성능 평가

3. 결 론

본 논문에서는 제안 한 CAV 알고리즘은 웹 서버에 접속한 세션이 로봇인지 일반 사용자를 실시간으로 구분하기 때문에 서버에 악영향을 끼칠 로봇을 실시간으로 차단 할 수 있다. 본 논문에서는 알고리즘의 신뢰도를 검증하기 위해 성능을 평가하였다.

CAV 알고리즘의 성능은 일반 사용자를 일반 사용자로 분류하는데 79~87%의 정확도를 보였으며, 로봇을 로봇으로 분류하는 경우는 95~99%의 정확도를 보인다. 또한 본 알고리즘의 검증을 위해 선별 된 11개의 로봇들의 경우 90%이상의 탐지율을 보인다. 여기서 11개의 로봇은 Microsoft, Goolgle, Yahoo, HTTPClient, IndyLibrary, IVM, CFNetwork, Libwww, GNU, Girafa 그리고 Powermarks 이며, 상업적 혹은 교육적인 용도로 널리 쓰이는 로봇들이다.

향후 연구는 로봇과는 비슷하지만 또 다른 주제인 로봇 이외의 비정상 프로그램의 검출이다. 로봇 이외의 비정상 프로그램 또한 직접 접속하는 일반 사용자와의 다른 차이점이 존재 할 것이기 때문에 이러한 특성을 연구하여 서버에 해악을 끼치는 비정상 프로그램에 대한 검출을 시도할 것이다.

참고 문헌

[1] P.-N. Tan and V. Kumar. Discovery of web robot sessions based on their navigational patterns. Data Mining and Knowledge Discovery, 6:9. 35, 2002.

[2] D. Lee. Web Robot Detection on Real-Time using a Composite Attribute Vector. Master’ s thesis, Korea advanced institute of science and technology, 2009

[3] J. Lee, Metrics for Classification of Web Robots: An Empirical Study on Over One Billion Requests, Master’ s thesis, Korea advanced institute of science and technology, 2008