Anti-Crawling Teknikleri

Bedirhan Urgun, Aralık 2010, WGT E-Dergi 7. Sayı

Yazıya başlamadan önce bu konuda ki terminolojiden bahsedecek olursak;

Deep Linking

Linklerin ana sayfa yerine spesifik bir sayfayı veya dokümanı göstermesidir.

Örnek; http://www.webguvenligi.org/wp-content/uploads/2008/11/link_kesfetme.pdf

Web Indexing

Web sitelerinin içeriği hakkında arama motolarının kolaylıkla servis verebilecekleri metadata ve terimler oluşturma tekniğidir.

Web Scraping

Web sitelerinden bilgi çalınması tekniğidir. Web Indexing'den farklı olarak Scraping'de amaç daha çok kopyalanan içeriğin saklanarak analiz edilmesidir. Analiz işlemi çok komplex olabileceği gibi olduğu gibi kullanılma şeklinde de olabilir. Web otomasyonu ile de özdeşleştirilebilen scraping, online ücret karşılaştırması, hava durumu denetimi, web sitesi değişiklik yönetimi gibi konularda da kullanılabilmektedir.

Hot Linking / Inline Linking / Leeching / Piggy-Backing / ...

Ana bir sitede link olarak kullanılan bir nesnenin (imaj v.b.) başka bir site içerisinde kullanılmasıdır. Bu şekilde ikinci site, tam olarak görüntülenmeyen ilk sitenin bandwith'ini arttırmaktadır.

Referer HTTP başlığının kullanılması Hot Linking'e karşı uygulanan en önemli önleme tekniğidir.

Cloaking

Bir engelleme tekniği olarak algılanabilecek cloaking, ana sitenin normal kullanıcılara gösterilen içeriğin belirli kullanıcılara (istek bazında veya IP tabanlı olarak) özel olarak değiştirilmesidir.

Teknikler

Crawling yapan son kullanıcı agent'ları genellikle otomatik bot yazılımlardır. Bu yazılımlar standard HTTP programlama ile geliştirilmiş uygulamalar olabileceği gibi tarayıcı tabanlı DOM'dan anlayabilen daha gelişmiş uygulamalar da olabilir.

Pozitif User-Agent Listesi

HTTP başlıklarından User-Agent, web sitesine isteği yapan kullanıcının kullandığı uygulamanın hangi yazılım olduğunu ve versiyonunu belirtmektedir. Bazı crawlerlar bu başlığın değeri olarak standard piyasada bulunan User-Agent değerlerini kullanmayabilirler. Bu başlığın değerinin standard beklenen User-Agent'lara (Internet Explorer, Firefox, Iphone, Nokia, Safari, Opera, v.b.) göre kontrolü ile bir önleme tekniği olarak kullanılabilir.

Geçerli Session ID Kontrolü

Uygulamalar genellikle kullanıcıları otomatik izlemek için session id (çoğu zaman cookie) kullanırlar. Sitelere IP'lerden yapılan ilk istekler bu bilgiyi henüz elde etmemişlerdir. Ancak ikinci istekler geçerli bir cookie (session id) bilgisini içermek zorundadırlar. Bu bilgiyi göndermeyen kullanıcılara karşı yaptırımlar uygulanabilir.

Cookie bilgisini tutan session tabanlı crawler'lara karşı ise daha karmaşık, özel ve granüler anti-crawling teknikleri geliştirilebilir (Bknz: İleri Teknikler).

Görünmez Linkler

Otomatik crawler'lar tarafından scrape edilen sayfalar içindeki linklerin görünüp görünmemesi bu crawler'ların ilgili linkleri isteme davranışlarını belirleme açısından kullanılabilir. Örnek;

```
<a style="display:none" href=http://www.webguvenligi.org/dontfollow.php />
```

Bu linki isteyen bir son kullanıcı ya uygulamayı analiz veya siteyi crawl etmektedir.

Javascript Tuzakları

Başlangıç, orta ve çoğu bazı ileri seviye crawlerların javascript destekleri yok veya gelişmemiştir. Sayfa içerisinde kullanılabilecek javascript tuzakları bu crawler'ların belirlenmesinde önemli rol oynayabilir. Browser tabanlı crawler'ların bile bazı durumlarda javascript/ajax tabanlı linkleri bulmaları ve çalıştırmaları kısıtlı olabilir.

http://code.google.com/p/wivet/wiki/CurrentResults

Örnek:

```
...
<html>
 <script type="text/javascript" >
  $(document).ready(function(){
   $("#link").each(function(){this.href = "youarenotcrawler.php";});
  });
 </script>
 <body  class="body">
   <a id="link" href="" target="body">click me</a>
 </body>
</html>
```

Ana sitenin javascipt desteği olmadan kullanıcılarda çalışmaması bu önlemin kullanılabilmesinde en önemli avantajdır.

Flash / Java Tuzakları

Javascript tuzaklarının aşılmasında sonra uygulanabilecek en önemli ve komplex tuzak Flash ve Java uygulamaları ile oluşturulabilecek otomatik linklerdir.

IP / İstek Sayısı Kısıtlama

Hem hizmet dışı bırakma (Denial of Service - DoS) hem de crawling önleme tekniklerinden biri de IP veya Session / İstek sayısı kısıtlamasıdır.

Belirli bir IP'den veya Session'dan gelen istek sayısı dakika ve saat üzerinden belirlenen bir maximum sayıyı aştığında bu IP'lere veya Session'ı kullanan IP'lere yaptırım uygulanabilir.

İleri Teknikler

Session Tabanlı İmzalar

Uygulamalarda son kullanıcılar için oturum (session) kullanımı zorunlu hale getirilmişse, yani aksi halde uygulamalar çalışmıyorsa, bu kısıt daha ileri seviye izleme için kullanılabilir.

Aynı cookie bilgisi ile gelen isteklerin bazı HTTP başlıkları kullanılarak imzaları üretilebilir. Uygulama tarafında ise ilgili oturumda istek bazlı kontroller daha granüler bir şekilde gerçekleştirilip yaptırımlar uygulanabilir. (Bazı durumlarda proxy arkasındaki normal kullanıcılar crawler'lardan ayrılabilir)

Kısa süre içinde oturum (cookie) değiştiren IP'ler ise belirli bir maximum değer aşıldığında yaptırıma tabi tutulabilir.