Benemérita Universidad
Autónoma de San Luis
Potosí (UASLP), México

Conferencia internacional
BIREDIAL-ISTEC
17-18-19 OCTUBRE 2016
congresos.unlp.edu.ar/**biredial-istec**

Detección de bots en reportes estadísticos

Juan Manuel Catá
Ariel J. Lira
Marisa R. De Giusti

PREBI prebi.unlp.edu.ar
SEDICI sedici.unlp.edu.ar
UNIVERSIDAD NACIONAL DE LA PLATA

El problema

Los repositorios digitales concentran una gran cantidad de enlaces entrantes y muchos contenidos de calidad por lo que resultan de mucho interés para los bots que navegan la World Wide Web

¿Qué es un bot?

Un bot es un software autónomo capaz de llevar a cabo tareas concretas e imitar el comportamiento humano.

Objetivos típicos de los bots

Aceptables

Descarga e indexación de páginas y recursos: buscadores

Análisis de links: visibilidad, SEO

Descarga de archivos: para análisis de las publicaciones, detección de plagio, etc.

No aceptables

Buscan acceder a información protegida

Encontrar vulnerabilidades

publicar datos no autorizados

Ads

Comportamiento de bots

Correcto

se identifican como bots en el userAgent - *(Mozilla/5.0 (compatible; robot de Google/2.1; +http://www.google.com/bot.html))*

respetan robots.txt

no realizan más de 1 acceso en simultáneo

Consecuencias

Permiten difundir los contenidos que el sitio desea

aumenta el impacto

Comportamiento de bots

Incorrecto

no cumplen con las reglas mencionadas

Consecuencias

degradan la performance

ensucian las estadísticas

Ejemplos

Bots conocidos

- Contenidos
  - Googlebot, Bing, Baidu, Yahoo Slurp.
- Visibilidad
  - Majestic SEO, Ahrefs, entre otros
- Otras aplicaciones específicas

Bots desconocidos

Accesos de bots

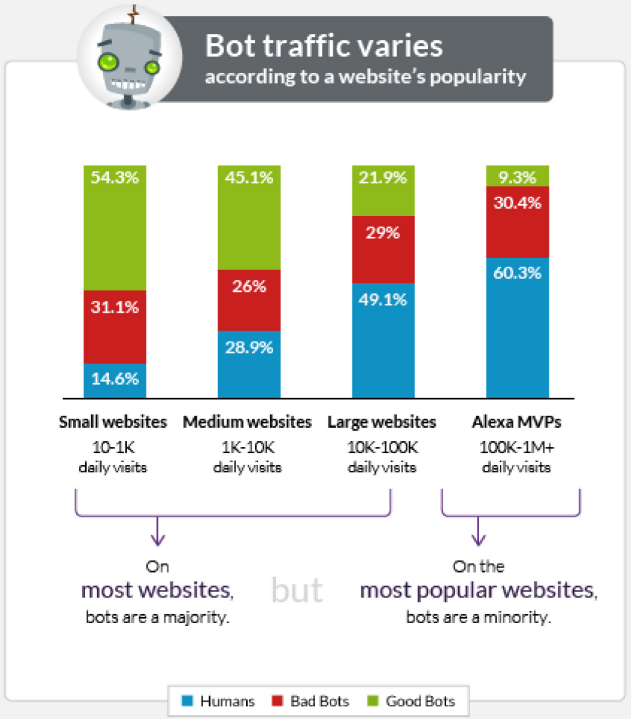


Gráfico extraído del artículo “2015 Bot Traffic Report: Humans Take Back the Web, Bad Bots Not Giving Any Ground”

https://www.incapsula.com/blog/bot-traffic-report-2015.html

Influencia de los bots en las estadísticas

Los accesos al sistema de bots no identificados generan cientos y miles de registros de accesos espurios que llevan a estadísticas poco confiables.

Si los datos estadísticos no son confiables, se pierde el potencial que tienen las estadísticas del repositorio como herramienta de gestión

Estadísticas de un repositorio

Cantidad de accesos a un recurso

Documentos más descargados

Acceso a los recursos de una colección

Distribución de accesos por país de origen

Tasa de accesos por fecha

Fechas con mayor cantidad de accesos

entre otras

Estadísticas en DSpace

DSpace incluye, desde la versión 1.6, un módulo de estadísticas:

registrar todos los accesos al sistema, tanto de usuarios normales como de bots

basado en Apache Solr


```
"start": 0,
"docs": [
  {
    "ip": "190.221.183.227",
    "referrer": "http://www.google.com.ar/url?s
    "dns": "host227.190-221-183.unsa.edu.ar.",
    "userAgent": "Mozilla/5.0 (Windows NT 6.1;
    "continent": "SA",
    "countryCode": "AR",
    "city": "Buenos Aires",
    "latitude": -34.603302,
    "longitude": -58.3816,
    "isBot": false,
    "id": 103,
    "type": 0,
```

Detección de bots en DSpace

Actualmente DSpace cuenta con mecanismos para identificar los bots que acceden al repositorio

- Mediante listas de ip
  - limitado
  - desactualizado
- Mediante listas de user-agent
  - Hay bots que no se identifican a través del user agent

Solución propuesta

Desarrollar una herramienta que detecte patrones de comportamiento de bots a partir del análisis de los accesos registrados y que permita:

- filtrar accesos existentes de sistemas automáticos
- Mantener la información obtenida para considerar los accesos futuros

La herramienta debe:

- implementar una heurística de detección basada en reglas
- ser configurable
- integrable a DSpace

Reglas

Estas reglas analizan los registros de acceso en busca de patrones de comportamiento sospechosos. Cada una de estas reglas implementan un criterio específico de búsqueda.

Algunos ejemplos:

- Cantidad de accesos durante un período de tiempo
- Análisis de agentes de usuarios
- direcciones IP pertenecientes a la misma subred

Diseño modular

La ventaja de utilizar este modelo es la adaptabilidad y extensibilidad que provee

- Agregar o quitar reglas del conjunto de reglas que serán ejecutadas
- Crear e integrar nuevas reglas que se adapten a las necesidades del repositorio

Configurable

No todos los repositorios son iguales:

- Volúmenes distintos de información
- Flujos distintos de la información
- Accesos de distintos tipos de bots

Por lo que la definición de la heurística debe ser un proceso extremadamente dinámico que permita ajustar los parámetros de búsqueda de las reglas.

Integrado a DSpace

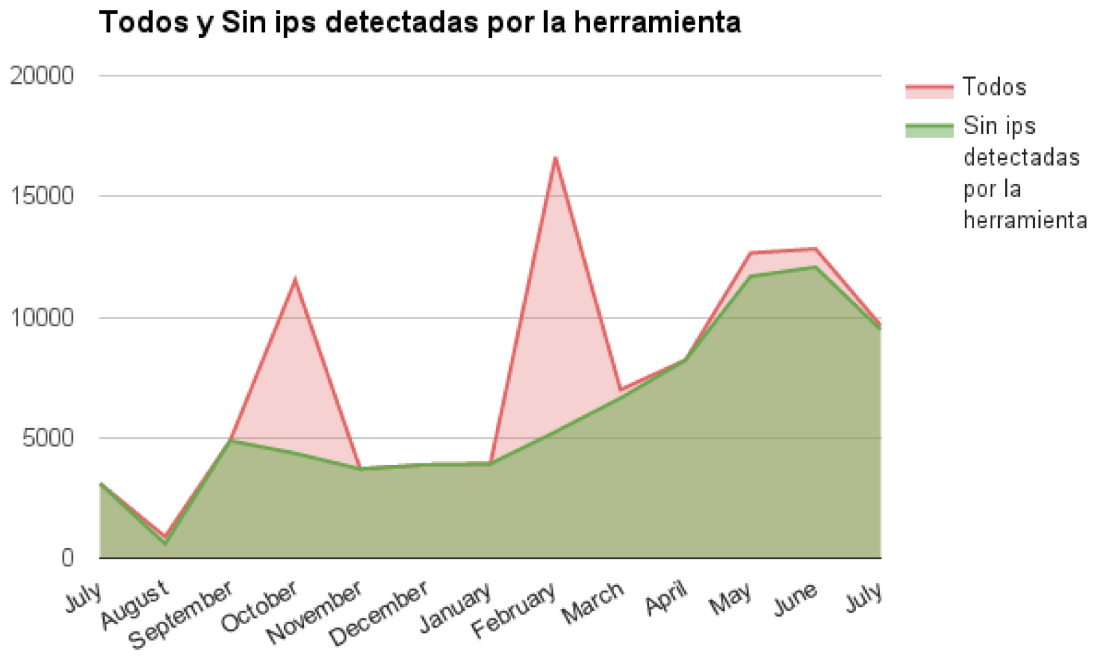
El prototipo de la herramienta funciona como una extensión del módulo de gestión de estadísticas que provee DSpace, por lo que si es aceptado por la comunidad de desarrollo, estaría disponible en cualquier instalación de DSpace, esto, sumado a la capacidad de configuración, lo convierten en una herramienta de suma utilidad para cualquier repositorio implementado con DSpace

Resultados preliminares

Se realizaron pruebas preliminares sobre el repositorio CIC-digital de la comisión nacional de investigaciones científicas de la provincia de Buenos Aires (Argentina)

1. Se recopilaron los datos de accesos a ítems del repositorio entre julio del 2015 y julio del 2016
2. Se ejecutó la herramienta buscando IP's que hayan realizado más de 50 accesos a ítems por hora, en ese período.
3. Posteriormente, las IP's detectadas fueron marcadas como bot y se volvieron a recopilar los datos del paso 1 filtrando las IP detectadas en el paso 2

Impacto en las estadísticas



https://docs.google.com/spreadsheets/d/18yd6OW3iGx_IhlWrUm9NeOQVHy19p7P-9uUeNItu7NY/edit#gid=0

Conclusiones y trabajos futuros

Los resultados preliminares convalidan el sentido del desarrollo y obligan a continuar con el refinamiento de la herramienta, particularmente para reducir la cantidad de falsos negativos que aún escapan al detector y hacer un análisis más exhaustivo de los resultados obtenidos para prevenir falsos positivos.

A futuro:

- más y mejores reglas
- bloqueo automático de bots maliciosos a partir de otras herramientas como fail2ban
- mejores reportes

Juan Manuel Cata, juan@sedici.unlp.edu.ar
Ariel J. Lira, alira@sedici.unlp.edu.ar
Marisa R. De Giusti, marisa.degiusti@sedici.unlp.edu.ar