Detección de bots en reportes estadísticos

Catá, Juan Manuel[1]; Lira, Ariel Jorge[2]; De Giusti, Marisa Raquel[3]

Resumen

Las estadísticas de un repositorio institucional son una herramienta básica que asiste el proceso de toma de decisiones y gestión del repositorio. Por este motivo, es importante que la información provista por estas estadísticas sea información precisa y confiable, en particular los registros de acceso y descarga.

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. Si bien la mayoría de los bots respetan las reglas básicas establecidas en los archivos robots.txt, muchos de ellos no lo hacen e incluso hay algunos que no se identifican como tales y se hacen pasar por agentes de usuario normales. A pesar de las medidas que se toman para evitar el acceso de bots maliciosos, un número importante de estos logra filtrarse y efectuar miles de accesos indeseados. Se genera, en consecuencia, gran cantidad de datos espurios que llevan a estadísticas poco fiables y que en última instancia entorpecen el proceso de gestión del repositorio

Para solucionar el problema planteado, se comenzó desarrollar una mecanismo que, a partir del análisis, permita filtrar los accesos de bots normales y bloquear los accesos de bots maliciosos o con mal comportamiento.

Las pruebas iniciales con la herramienta han permitido identificar un número elevado de accesos correspondientes a bots maliciosos que, al filtrarlos, permiten obtener resultados estadísticos mucho más veraces.

Abstract

1 Estudiante avanzado de Licenciatura en Sistemas. PREBI-SEDICI, Universidad Nacional de La Plata.
juan@sedici.unlp.edu.ar

2 Licenciado en Sistemas. PREBI-SEDICI, Universidad Nacional de La Plata; CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. alira@sedici.unlp.edu.ar

3 Doctor en Ciencias Informáticas. Investigador independiente de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires; PREBI-SEDICI, Universidad Nacional de La Plata; CESGI,
marisa.degiusti@sedici.unlp.edu.ar

Statistics are an essential tool for institutional repositories which assists the decision making process and repository management. Therefore, the information they provide must be precise and reliable, specially those based on access and download logs.

When a digital repository grows and brings together large amounts of incoming links and high quality content, it acquires of great significance for bots. Most bots follow the basic rules established in robots.txt files, nevertheless many do not do it, and some of them do not identify themselves as bot masquerading as normal users. Despite the measures taken to avoid access to malicious bots, a large amount of them manage to seep and make thousands of unwanted access. Therefore a large number of spurious data is generated which leads to unreliable statistics and hinders the repository management process.

In order to solve this problem, a mechanism was developed to analyse, detect and filter access from malicious or misbehave bots.

Initial tests with this tool allowed to identify a large number of access comming from malicious bots that, after being filtered, allows to get much more sound and reliable statistics results.

Introducción

Las estadísticas de un repositorio institucional son una herramienta indispensable para la obtención de métricas que definen el alcance y el impacto del repositorio y de cada recurso dentro del mismo. Algunos reportes típicos podrían ser:

- Cantidad de accesos a un recurso o colección
- Documentos más descargados
- Accesos de los recursos de una colección
- Distribución de accesos por origen (continente/país/ciudad)
- Tasa de accesos por fecha
- Fechas con mayor cantidad de accesos
- entre otros

Estas métricas influencian la gestión y la toma de decisiones en los repositorios, por tal motivo es de suma importancia que los datos obtenidos sean fiables y precisos.

DSpace incluye un módulo estadísticas (Estadísticas DSpace, 2016) que ofrece reportes simples a nivel global, de comunidad, de colección y de ítem de:

- ranking de cantidad de accesos por país y ciudad
- cantidad total de accesos/descargas
- ranking de ítems más accedidos

El módulo está basado en el software de indexación de texto Apache Solr que registra todos los eventos vinculados al acceso al repositorio, a la actividad interna de su flujo de trabajo y al uso del módulo de búsqueda. Los datos más importantes que se asocian con los eventos de acceso y que se utilizan para el presente trabajo son:

- type: tipo de recurso accedido (item, comunidad, colección, etc)
- id: combinado con el type identifica el recurso.
- ip: IP desde donde proviene el acceso
- uid: id interno del objeto de DSpace (a partir de DSpace 6)
- continent: continente desde el cual se accede
- country: país desde el cual se accede
- city: ciudad desde donde se accede
- userAgent: agente de usuario utilizado al momento de acceder
- isBot: identifica si la ip pertenece a un bot o no
- referrer: identifica la dirección de la página que creó el vínculo con el recurso que está siendo solicitado
- además de muchos otros campos complementarios que aún no se utilizan como parte del análisis.

Este módulo registra todos los accesos, tanto los provenientes de bots como de usuarios normales, y utiliza un una etiqueta *isBot* para diferenciar los accesos de cada caso. El etiquetado de los accesos de bots se realiza de dos formas: 1) a partir de una lista de direcciones IP de bots conocidas y 2) a partir del header HTTP *userAgent* que proviene en la solicitud.

Sin embargo, a pesar de disponer de estas formas de detectar y filtrar el acceso de bots, diariamente acceden al sistema bots malintencionados (spiders, miners, crawlers, etc), que no se identifican como tales, y buscan acceder y descargar todo el contenido posible del repositorio, generando un número muy alto de accesos y contaminando los registros estadísticos.

Entre otros, se identifican 4 grandes grupos de bots que fueron tenidos en cuenta durante el desarrollo del prototipo:

1. Bots normales conocidos, como el de Google (GoogleBot), que respeta las reglas especificadas en el robots.txt. El módulo de estadísticas de DSpace ya incluye la funcionalidad para detectar y descartar estos casos a partir del procedimiento mencionado.
2. Bots normales aunque desconocidos: son bots que se identifican como tales, por ejemplo, mediante el user agent, y respetan el archivo robots.txt pero que no son conocidos en el sistema porque son nuevos o muy específicos.
3. Bots con comportamiento incorrecto: son procesos que buscan acceder y descargar todo o parte del contenido del repositorio sin respetar las reglas básicas establecidas en el archivo robots.txt. Este tipo de bots son muy perjudiciales porque como suelen realizar numerosos accesos concurrentes, suelen degradar la performance del sistema y hasta pueden provocar su caída.
4. Bots maliciosos: típicamente spiders o miners que buscan vulnerar el sistema para conseguir información protegida, publicar datos no autorizados, publicidad, o simplemente afectar negativamente el repositorio.

Propuesta

A partir de la problemática planteada y de la funcionalidad provista por Apache Solr y DSpace, se inició el desarrollo de una herramienta modular y configurable, integrada al módulo de manejo de estadísticas de DSpace (stats-util) que, mediante el análisis de los eventos de accesos registrados en el módulo de estadísticas, intenta:

1. detectar direcciones IP con comportamientos anómalos característicos de bots no identificados y marcar los eventos asociados con dicha dirección con el flag isBot
2. Guardar la dirección detectada en una lista de direcciones IP de bots conocidos para que sean detectados de forma temprana en futuros accesos.(AUTOR, 2008)

Un esquema de trabajo similar se plantea en *(*Stassopoulo A. y Dikaiakos M. D *2006)"* en el cual se propone un mecanismo para evaluar los registros de accesos o *logs* provistos por servidores web para detectar la actividad de *crawlers*.

El proceso de detección ejecuta una serie de *Reglas*, cada una de las cuales busca detectar patrones de comportamiento anómalos específicos y, de acuerdo con la probabilidad calculada de que las direcciones IP detectadas sean bots, marcarlas automáticamente o bien generar un reporte que pueda ser analizado posteriormente por un humano para definir si realmente se trata de un bot, o si es un caso aislado y particular.

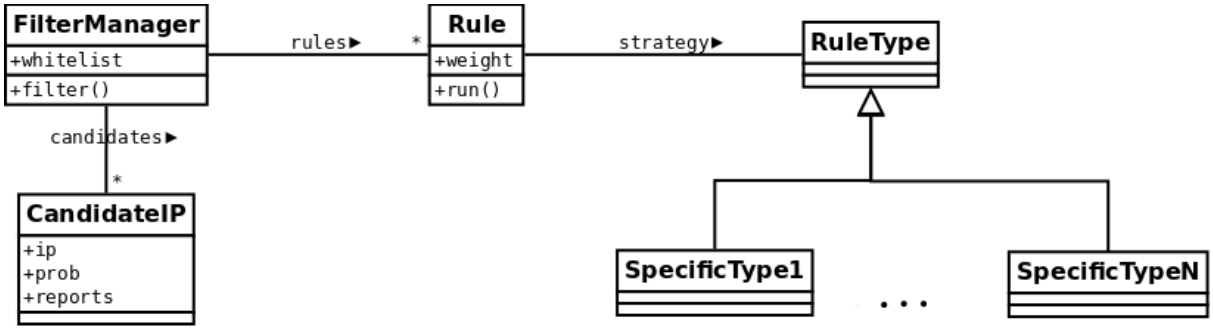
La funcionalidad de la herramienta no es filtrar ni bloquear el acceso de bots al sistema, sino mantener los registros estadísticos sanos e identificar cuánto de lo que se registró corresponde a accesos de bots, para que las próximas métricas que se realicen basadas en estas estadísticas, sean lo más correctas posibles.

Este trabajo se está llevando a cabo sobre CIC-Digital, el repositorio institucional de la Comisión de Investigaciones Científicas (CIC) de la provincia de Buenos Aires (Argentina), y se prevé su aplicación luego en el repositorio institucional de la UNLP, SEDICI.

Implementación

La herramienta consta de un controlador central, encargado de ejecutar las reglas especificadas en un archivo de configuración, y mantiene una lista con las direcciones IP candidatas junto con la probabilidad de que dicha dirección IP sea un bot. Una vez ejecutadas todas las reglas, el controlador informa los resultados y marca las direcciones IP detectadas como bots en los registros estadísticos.

El modelo es ligeramente más complejo, pero podría resumirse de la siguiente manera



Las reglas son elementos configurables encargados de analizar los eventos de acceso del módulo de estadísticas, para evaluar la actividad de cada dirección IP y en caso de detectar una dirección IP sospechosa, se le asigna a esta una “probabilidad de bot” de acuerdo a cuán seguro se esté de que dicha IP sea un bot. Por ejemplo, si se detecta que una dirección IP tuvo 100 accesos en un mes, la probabilidad asociada es baja, en cambio si tuvo mas de 1000 accesos, la probabilidad es mucho mayor (los valores usados son a modo de ejemplo, cada repositorio debería manejar los valores correspondientes a sus propias necesidades).

Tipos de reglas

El tipo de regla (*RuleType*) contiene la implementación de la lógica de búsqueda que tendrá una regla. La herramienta permite definir varias reglas con un mismo mismo *RuleType* a las cuales se le especifican parámetros diferentes que se adecúen al caso buscado, por ejemplo: un *RuleType* que busca cantidad de accesos[4] en un período de tiempo, podría ejecutarse una vez para analizar los accesos a ítems, y otra para la cantidad de descargas ese período de tiempo.

Una de las ventajas más importantes de la utilización de estas reglas como módulos individuales es la adaptabilidad y la extensibilidad. A futuro podrían crearse nuevas estrategias de búsqueda que se adapten mejor al repositorio que se está analizando.

Uno de los tipos de reglas creados hasta el momento se encarga de evaluar la cantidad de descargas (o accesos) en períodos fijos de tiempo. La regla es completamente configurable y permite ajustar el lapso de tiempo, la cantidad de accesos mínimos requeridos para considerar una dirección IP sospechosa, el tipo de acceso (bitstream, item, collection, etc), entre otros ajustes. Este tipo de regla permite detectar secuencias de acceso que sería imposible de realizar para un humano.

Con este criterio se detectaron por ejemplo muchos casos de bots no identificados que realizan:

4 *Los accesos se diferencian por tipo, (item, bundle, collection, etc). Cuando se habla de cantidad de accesos podría ser cualquiera de estos tipos*

- cientos de descargas por hora y absorben todos los archivos PDF del sitio
- miles de accesos continuos períodos largos como un mes

También se definió otro tipo de regla que realiza un análisis por subred a partir de la identificación de accesos repetidos desde direcciones de IP con raíz similar. Si se detectan muchos accesos provenientes de una misma subred podría tratarse de una botnet y debe ser tenido en cuenta.

Configuración

La definición de la heurística de diagnóstico es un proceso extremadamente dinámico dado que hay varios factores que requieren ajustar, agregar o eliminar reglas de forma permanente. Algunos factores son:

- se debe realizar un ajuste de los parámetros de cada regla
- se pretende usar el desarrollo en varios repositorios y no todos los repositorios reciben los mismos bots
- el comportamiento de los bots puede ser cambiante y esporádico
- la ejecución de algunas reglas muy complejas pueden afectar la performance del sistema

Por tal motivo, es necesario que la herramienta en general y las reglas en particular se adapten para cubrir las necesidades particulares de cada caso. A continuación se detallan algunas configuraciones posibles:

- Las cadena de reglas que serán ejecutadas en orden
- La configuración específica de cada regla: se indica el peso de la misma (valor que determina la probabilidad de bot) y parámetros complementarios que dependen directamente del tipo de regla a ejecutar. Por ejemplo, una regla de cantidad de accesos por mes tendrá los siguientes parámetros propios definidos:
  - Mínimo de accesos para considerar la dirección de IP como sospechosa
  - Tipo de recurso a considerar
  - Período de tiempo en el cual se evalúa la condición
- Una lista segura o *whitelist* de direcciones IP que no deben ser tenidas en cuenta al momento de analizar los registros porque es garantizado que son confiables y no son bots. Pueden existir casos de direcciones IP con un comportamiento inválido para un usuario normal, pero que no obstante eso de antemano se conoce que no son bot como por ejemplo: la subred institucional desde la que los administradores ingresan cientos de veces por día, una ip de una red NAT, entre otros.

Caso de prueba

En primer lugar se ejecutó la herramienta sobre el repositorio CIC-Digital analizando la cantidad de accesos a items por hora que se realizaron entre julio de 2015 y julio del 2016, tomando como criterios de búsqueda aquellas direcciones IP que accedieron a más de 50 ítems por 1 hora. Se detectaron direcciones IP que efectuaron desde 60 accesos en una hora, otras con más de 300, e incluso un caso con más de 1000 accesos en 1 hora.

También, como se menciona en (Balla et Al, 2011) la hora a la que se realizan dichos accesos puede ser un factor importante, en el artículo se habla de *night requests* y se refiere a los accesos realizados entre las 2am y las 8am.

En segundo lugar se realizaron consultas sobre los registros de acceso del módulo de estadísticas para generar un gráfico filtrando según 2 criterios, a saber:

1. Todos los accesos realizados entre el 01/07/2015 y 01/07/2016
2. Todos los accesos en el mismo período sin considerar los accesos de las IP sospechosas detectadas previamente con la herramienta.

Como se muestra en el gráfico a continuación, cuando se consulta por todos los accesos realizados al sistema se ven picos de accesos sospechosamente altos en los meses de octubre del 2015 y febrero del 2016 teniendo en cuenta los accesos que se venían registrando. Luego de omitir las direcciones IP detectadas por la herramienta, los resultados obtenidos cambiaron drásticamente, se eliminaron los picos de octubre y febrero, y se aprecia una leve mejora en los meses de agosto del 2015, marzo, mayo y junio del 2016.

Conclusiones

Las pruebas preliminares de la herramienta permiten observar numerosos casos de direcciones IP pertenecientes a bots no identificados que acceden a items y descargan contenido del repositorio, los cuales registran números muy elevados de accesos, contaminando seriamente los registros estadísticos. Luego de filtrar dichas direcciones IP se logró conseguir estadísticas mucho más fiables.

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.

La flexibilidad de configuración y el beneficio en la calidad de estadísticas resultantes hacen de éste un producto muy útil para cualquier repositorio basado en DSpace. Por tal motivo, se prevé proponer el desarrollo para que sea integrado a DSpace en un futuro cercano.

Bibliografía

Stassopoulo A. y Dikaiakos M. D. (2006) Crawler Detection: A Bayesian Approach. Proc. Int'l Conference on Internet Surveillance and Protection (ICISP'06), 16-21. doi:10.1109/ICISP.2006.7.
http://linc.ucy.ac.cy/publications/pdfs/2006-CISP-CrawlerDetection.pdf

Balla, A.; Stassopoulou, A.; Dikaiakos, M. D. (2011) Real-time Web crawler detection. Proc. of 18th International Conference on Telecommunications (ICT'11), pp. 428-432. doi:10.1109/CTS.2011.5898963. http://linc.ucy.ac.cy/publications/pdfs/2011-ICT-ANapa.pdf

Modulo de Estadísticas DSpace - https://wiki.duraspace.org/display/DSDOC5x/SOLR+Statistics Revisado el 8 de agosto de 2016.

CIC-digital - http://digital.cic.gba.gob.ar (revisado al 9/8/2016)