Proyecto Fin de Carrera

Detector y corrector automático de ediciones maliciosas en Wikipedia

Autor: Emilio José Rodríguez Posada
Director: Manuel Palomo Duarte

Marzo de 2009

Ingeniería Técnica en Informática de Sistemas
Universidad de Cádiz

Índice

- Introducción (5 minutos)
  - Wikipedia
  - La problemática de los vandalismos
- AVBOT (10 minutos)
  - Desarrollo
  - Funcionamiento
- Demostración (5 minutos)
- Conclusiones y trabajos futuros (5 minutos)

Wikipedia, ¿qué es?

- **Wikipedia** es un proyecto fundado en 2001 para crear una **enciclopedia libre**
- Ha tenido un crecimiento vertiginoso:
  - 15 millones de artículos en 8 años
- Comunidad **global**:
  - Más de 250 idiomas
- Penetración:
  - Unas 4000 páginas servidas por segundo

Wikipedia, ¿cómo funciona?

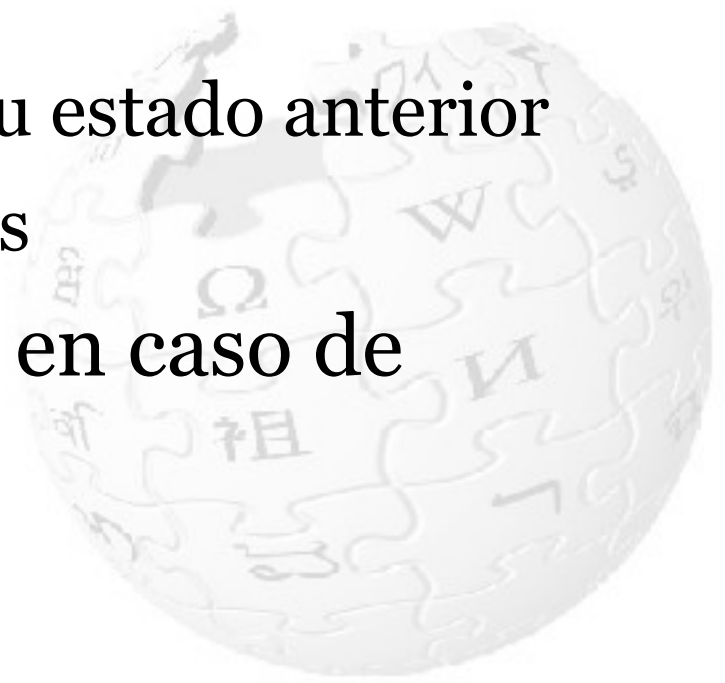
- Utiliza el software **MediaWiki** (software libre)
- **Cualquiera** puede modificar los artículos:
  - Ventajas: entorno de colaboración extrema
  - Inconvenientes: sensible a ataques:
    - Vandalismos (palabras soeces, insultos…)
      - X político fue un…
    - Blanqueos (eliminación de contenido válido)
      - Eliminar todas las Copas de Europa al Real Madrid
    - Pruebas de edición (pruebas de novatos)
      - En cualquier sitio: “Hola, soy Pepito, probandooooo”
  - Estos ataques son fácilmente reparables, aunque es una **labor tediosa** de realizar

Objetivos

- Se pretende crear un **programa que proteja Wikipedia** en español de estas ediciones maliciosas:
  - Corrección automática de vandalismos repetitivos
    - El programa nunca podrá sustituir a un humano
  - Evita que los usuarios pierdan su valioso tiempo
  - Avisar de vandalismo reincidente
- Cambios acordes a manual de estilo en artículos nuevos

¿Cómo se hace?

- Con un *bot* vigilando las 24/7
- **Capturando** las modificaciones por IRC a modo de RSS
- **Analizando** el texto anterior y posterior al cambio:
  - Se obtiene el texto modificado y su estado anterior
  - Se analizan los cambios entre ellos
- **Restaurando** el texto anterior en caso de vandalismo



¿Cómo detecta vandalismos?

- Mediante una serie de **expresiones regulares** que recorren el texto nuevo. Ejemplos:
  - idiota/s: (?i)i+d+i+o+t+a+s* (-2 puntos)
  - chingar: (?i)c+h+i+n+g+a+r+ (-3 puntos)
- Cada expresión tiene un **peso**. Se va acumulando
- Cuanto menor sea el total acumulado más probable será que el *bot* revierta

¿Cómo detecta blanqueos?

- Calculando el **porcentaje de texto retirado**
- Suponiendo que el texto anterior tiene 1000 bytes:
  - Texto nuevo = 600 bytes (Cambio legítimo, no se actua)
  - Texto nuevo = 34 bytes (Cambio extraño, posible blanqueo)
- El **porcentaje crítico** es 16% (Se activa al eliminar los 5/6 del total)

¿Cómo detecta pruebas de edición?

- Al igual que los vandalismos, con **expresiones regulares**
- Se envía un **aviso más amigable** que en el caso de vandalismo:
  - "Para pruebas utilice la Zona de pruebas habilitada para ello, por favor"

Cambios según manual de estilo

- Cada día se crean ~400 artículos
- Muchos de ellos no siguen normas de estilo
- Tampoco incluyen ***interwikis***:
  - Enlaces entre artículos homólogos en Wikipedias en otros idiomas.
- El *bot* unifica el estilo y trae de Wikipedia en inglés los *interwikis*

Avisos generados

- Si el *bot* deshace una modificación, envía un **mensaje** al usuario:
  - Vandalismos y blanqueos: Ruega que se detenga
  - Pruebas: Envía al novato a la "Zona de pruebas"
- Si el usuario insiste en vandalizar, se genera un **informe** que es enviado a los administradores

Dependencias y recursos

- Realizado exclusivamente con **software libre**:
  - **Python**
  - Biblioteca **irclib** para recoger la información publicada en IRC
  - Biblioteca **pywikipediabot** para interactuar con Wikipedia
- Conexión a Internet
- Recomendable: ejecución 24/7

Rendimiento

- **Evolución** de la detección de vandalismos (tras meses de pruebas):
  - Estático: únicos pesos válidos -1 ó +1
  - Puntuaciones: libertad total para pesos
  - Cálculo de densidad: toma de decisiones en función de la cantidad de texto nuevo
- Modificación de la lista de expresiones en **tiempo real**
  - La lista es pública

Dificultades superadas

- Gran volumen de ediciones a analizar, ~30.000 al día:
  - Solución: **Hilos** (threads)
- Gran número de expresiones regulares:
  - Solución: **Precompilado**. 15 seg -> 0,001 seg
- Reducción de falsos positivos:
  - Solución: Expresiones regulares con **pesos**

Demostración

- Ejecución en terminal
- Últimos vandalismos reparados
- Detalle de alguno de ellos

Desarrollo y pruebas

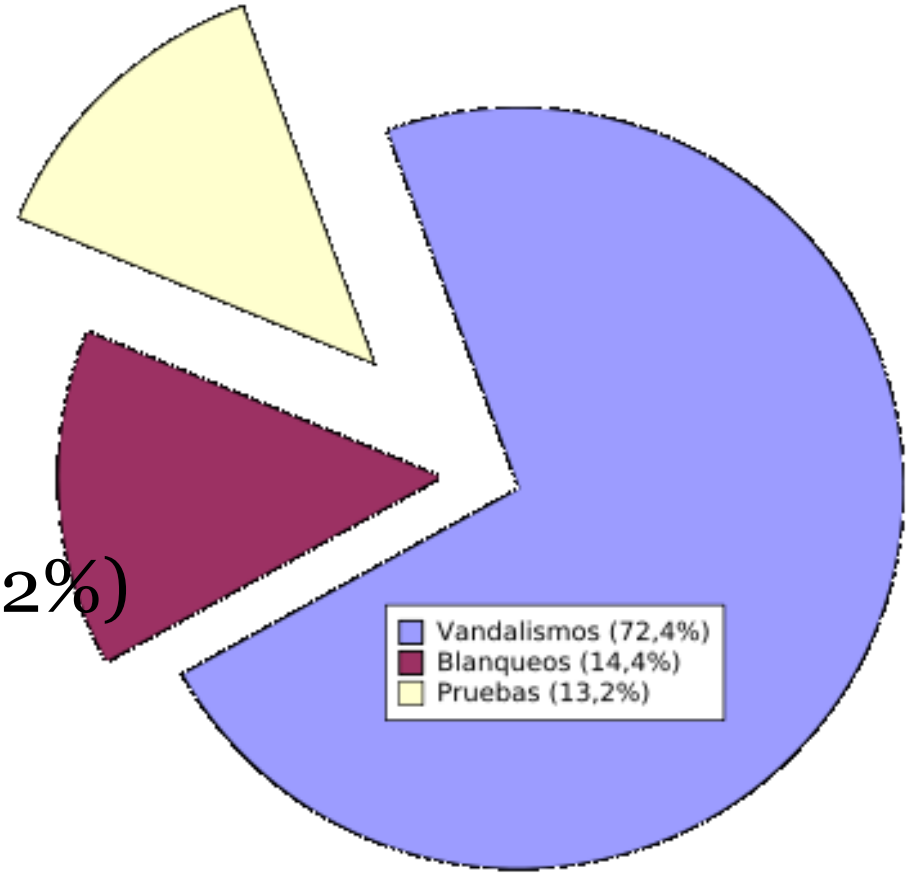
- Metodología *e**X**treme **P**rogramming*
- Desarrollo iterativo incremental
- La mejor validación que existe:
  - Cada vez que alguien modifica Wikipedia, pone a prueba al *bot*
- Continuas **mejoras sobre** los productos resultados de las **etapas anteriores**

Estadísticas

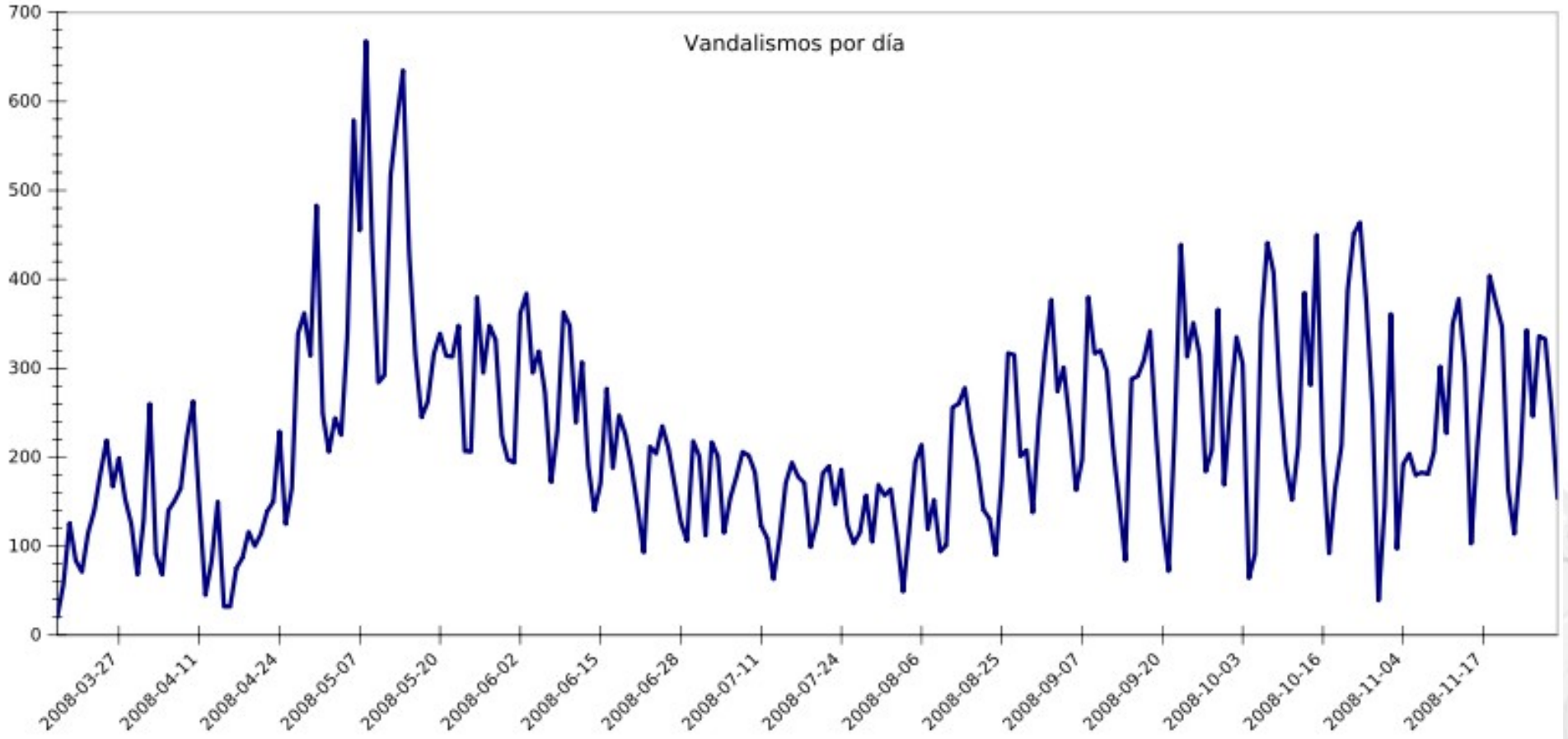
- Sobre el *bot*:
  - Casi **100.000** vandalismos revertidos
  - ~0,5% de error
  - Más de 5000 horas de funcionamiento
- Sobre vandalismo en general:
  - Gran parte vandalismo proviene de anónimos
  - Todos los países vandalizan por igual
  - Artículos muy vandalizados: Cataluña, Augusto Pinochet, Idioma valenciano, Cuba...

Estadísticas. Intervenciones (%)

- Porcentajes:
  - Vandalismos (72,4%)
  - Blanqueos (14,4%)
  - Pruebas de edición (13,2%)

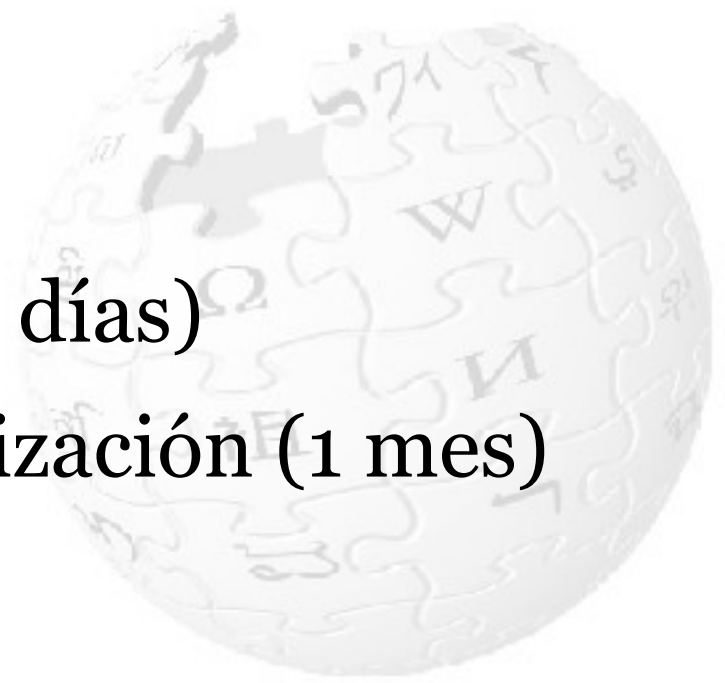


Estadísticas. Actividad en 2008



Calendario de trabajo

- Captura de datos (1 semana)
- Análisis de ediciones (Continuo, ~1 año)
  - Algoritmos (Continuo)
  - Lista de expresiones y pesos (Continuo)
- Auto-diagnóstico (3 días)
- Mensajes y avisos (2 semanas)
- Gestión de *logs* y estadísticas (5 días)
- Documentación e internacionalización (1 mes)



Conclusiones

- Alta **estabilidad**, +7 días de *uptime*
- Muy testeado, 5000 horas de ejecución
- **~100.000** vandalismos reparados
- Porcentaje de **error muy bajo**
- **Comunidad satisfecha**:
  - Otorgaron permiso para ejecutarlo continuamente
  - RoyFocker (administrador): "*...De verdad, que mil gracias por tanto trabajo que nos ahorra el bot...*"

Antes y después...

- Formación wikipédica desde 2005
  - Asuntos técnicos
- *Bot* como PFC desde mediados de 2008
- **Beca** "Conocimiento Abierto" en OSLUCA
- Participante en **Concurso Universitario** de Software Libre 2009

Futuro desarrollo

- Incorporar **nuevos patrones** de detección
- Controlar introducción masiva de texto
- Controlar **imágenes chocantes**:
  - Imagen de un pene en el artículo “Francia”
- Anti-traslados:
  - Renombrar “Selección española de fútbol” a “A por ellos oee oeee oeeeeee”
- Continuar **internacionalización**

Gracias por su atención
¿preguntas?