udp Escuela de Informática
y Telecomunicaciones
FACULTAD DE INGENIERÍA

# Predicción de patrones de comportamiento de usuarios en la web basado en web access log para un Admiinistrador de Contenidos Open Source

Jaime Guzmán
mail@jguzman.cl

Adin Ramirez* y Francisco Claude**

## 1. Antecedentes y Motivación

### 1.1. Contexto

La Web crece constantemente y por ende su infraestructura como también la concurrencia de los mismos sistemas, paralelamente se suman un costo exponencial de recursos que no son optimizados para poder dar un experiencia de usuario con calidad de servicio. Lo cual podemos entender que llegará un punto en que no tener servidores de gran rendimiento será lo óptimo para dar un calidad de servicio web, el ancho de banda de Internet no crecerá a la misma proporción. Adicionalmente las tecnologías para la creación de web dinámicas han evolucionado a favor del cliente, se tiene MEAN stacks que disminuyen considerablemente la carga de un servidor, por lo cual hoy en día un buen servicio web es proveer una balanceada carga dentro del cliente y el servidor.

Por lo mismo es de gran interés predecir lo siguientes movimientos que tendrá un usuario en un determinada web, entendiendo que la forma en que navega una persona es su comportamiento web, que se puede reflejar mediante Web Access Log. El registro de los mismo, de manera procesada ó pre-procesada ayudaría a ingenieros de desarrollo web y diseñadores, como a los mismo usuarios finales a tener una experiencia de usuario mejor.

Hoy en día las web no pueden ser simplemente dinámicas estas deben poseer una adaptabilidad a la demanda del usuario ó proveer información que permita adaptarse a los eventos, por lo cual es sumamente de interés profundizar este tópico.

### 1.2. Trabajos relacionados

En este tema participan dos áreas, por un lado existe trabajo para crear estructuras de eficiencias para predicciones basadas en algoritmos de compresión Claude [1] y por otro lado el uso de máquinas de aprendizaje para realizar clustering y predecir el comportamiento.

El tema de la predicción en la web en la literatura se ha presentado como un tema concurrente, abarcado por varios autores, tenemos los siguientes trabajos en orden cronológico:

1. Xing Dongshan And Shen Junyi [2], destacan que un modelo de Markov puede ayudar a predecir el comportamiento de un usuario, pero con ciertas limitaciones. Para solucionarlo presentan un nuevo modelo de Markov basado en un representacion de Tree Order Model, el cual es un híbrido entre un modelo de markov tradcional y una representación de Tree, bautizada como HTMM (Hybrid-Order Tree Markov Model). Su modelo fue presentado en 2002, da un importancia a conocer la predicción de los web access, dada la importancia de creación de redes, la minería de datos, e-commerce, y otras áreas.

2. Josep Domenech, Jose A. Gil, Julio Sahuquillo, Ana Pont [3], muestran un estudio de los rendimientos de técnicas de recuperación de datos, las misma que se pueden utilizar para dar una entrada ideal a algoritmos de aprendizaje ó algoritmos de predicción. Los conceptos mas importantes son las nuevas variables de caracterización que se le suman a la predicción propiamente las cuales son: temporalidad, espacio y geografía. Además de comenzar un trabajo más elaborado de como tomar una predicción, se

*Profesor guía
**Profesor comisión

introducen conceptos como Predicciones genéricas o específicas, variables de uso de recursos a nivel de red ó nivel procesamiento y finalmente lo que es totalmente significativo es que un modelo predictivo puede ayudar a disminuir la latencia entre la petición del cliente y la respuesta de la web, dando así un mejor rendimiento y QoS.

3. Zeljko Eremic, Dragica Radosav, Branko Markoski [9], casi cinco años después los sitios web son cada vez mas dinámicos y responden a eventos cada vez más adaptables a los usuarios. El trabajo de Zeljko-Dragica-Branko propone una optimización de la ruta de navegación de los sitios y estructura de la navegación del sitio.

4. Yuhua Chen, Xin Chen and Haoyi Chen [8], en su trabajo dan una nueva perspectiva enfocada a dar una clara recomendación a los usuarios basada en la misma propuesta de este proyecto, los access log. El primer analisis realizados por los autores cubre las reglas asociativas que requiere un sistema de recomendación, pero en las pruebas propiamente tales encuentran que el analisis de los patrones detectadados dan una representación clara como optimizar web y finalmente mediante sus pruebas logran una recomendación de calidad.

5. A. Rajimol and G. Raju [7], en este estudio se hace ya una minería en los patrones de los accesos web, el enfoque es usar los registros de de accesos para crear subsecuencias y realizar comparaciones. Basado en este estudio es bastante más claro que en la literatura y academia se presenta un interés para poder anticipar el patrón de comportamiento de la web.

6. Liu Kewen [4], en este trabajo ya es un análisis mas profundo del web usage minning, parte de lo importante de este trabajo es que después de minar los registros de accesos, se puede lograr que la "bad data" sea reducida.

7. Poornalatha G, Prakash S Raghavendra [6], esta presentación estable que se puede usar máquinas de aprendizaje para hacer predicción basadas en distintas entre cluster, estos autores al igual que Domenech-Gil-Sahuquillo-Pont[2005] y Dongshan-Junyi, comporante el objetivo de optimizar los recursos tanto en redes(disminución de latencia) y experiencia de usuario.

8. Francisco Claude, Gonzalo Navarro y Roberto Konow [1], finalmente se llega a un interés en particular, este trabajo presenta un estructura de representación eficiente que permite dar una representación de web access log y ofrece las operaciones básicas de WUM.

## 1.3. Motivación

La motivación de este proyecto de titulo es lograr una predicción mediante web access log, determinando que metodología usada es la más eficiente, como también un predicción del comportamiento del usuario en la web, usando plataformas experimentales como administradores de contenidos, con rutas url limpias predeterminadas.

## 1.4. Descripción de la solución

Se implementará un algoritmo Lempel Ziv 78 y un modelo tradicional de markov, con un dataset predefinido. Ambas implementaciones medirá el rendimiento de ambas con el fin de encontrar un taxonomía de predicciones favorables el set de datos.

## 1.5. Objetivo General

El objetivo general del proyecto de título es poder encontrar predicciones basadas en patrones encontrados en access web log, que permitan a un CMS Open Source Adaptarse a los solicitudes del cliente final.

## 1.6. Objetivo Específicos

A continuación se detallan los objetivos específicos:

1. Estudiar y describir el estado del arte respecto a las variantes existentes de Modelos Predictivos, describiendo limitaciones pro y contra entre áreas de estudio, como futura implementaciones y mejoras a la web.

2. Implementar un algoritmo de compresión para usarlo como un algoritmo de predicción.
3. Implementar un algoritmo basado en modelo Markoviano que permita entregar una predicción
4. Usar y preparar conjuntos de prueba para medir, clasificar las predicciones recuperadas de las pruebas experimentales.
5. Ejecutar pruebas de rendimiento usando nuestra implementación, y comparar con algoritmos expuestos en la literatura.
6. Analizar resultados obtenidos y mostrar el uso predictor .

## 2. Metodología de trabajo

Se investigar en detalle el funcionamiento del algoritmo, para luego generar una implementación del algoritmo aproximado. Durante el desarrollo del proyecto de título, se tomarán decisiones de implementación que permitan llevar la propuesta teórica a una implementación práctica.

Ya implementado el algoritmo, se realizarán pruebas un dataset de la literatura y se creará un módulo para el CMS Drupal que permita generar predicciones.

Estas pruebas serán luego ejecutadas en una máquina de prueba y recopilados los resultados para ser analizados. El Análisis comprenderá como se relacionan las áreas de IR, Machine Learning y Compresión de Algoritmos con el fin de lograr unificar la propuesta de solución.

## 3. Cronograma de actividades, hitos y entregables

| Fecha | Actividad |
|---|---|
| 31/07/2015 | Presentación anteproyecto (firmado por profesor guía y comisión). |
| 02/08/2015 | Entrega resultados anteproyectos. |
| 04/08/2015 | Entrega anteproyectos corregidos. |
| 09/08/2015 | Entrega resultados correcciones. |
| 25/09/2015 | Marco de trabajo. |
| 23/09/2015 | Primer prototipo de propuesta. |
| 13/10/2015 | Resultados parciales. |
| 27/10/2015 | Entrega Memoria Titulo I/II firmada por profesor guía. |
| 11/11/2015 | Fecha límite para que la comisión entregue correcciones. |
| 18/12/2015 | Fecha límite para que se realicen correcciones. |

## 4. Resultados esperados

Se espera crear un estudio completo de las áreas de interés y la implementación de ambas metodologías para ambos casos Además se espera entregar un listado de puntos de evaluación para implementar un algoritmo en el backend de un CMS que permita crear las predicciones del usuario basado en su comportamiento.

## Referencias

[1] Konow R. Claude F. and Navarro G. :. Efficient indexing and representation of web access logs. 2014.

[2] Xing Dongshan and Shen Junyi. A new markov model for web access prediction. *Computing In Science & Engineering*, 2002.

[3] Julio Sahuquillo Ana Pont Josep Domenech, Jose A. Gil. Web prefetching performance metrics, a survey. *Performance Evaluation, An Internation Journal*, 2005.

[4] Liu Kewen. Analysis of preprocessing methods for web usage data. *Intemational Conference on Measurement, Information and Control (MIC)*, 2012.

[5] Jia li. Research of analysis of user behavior based on web log. *International Conference on Computational and Information Sciences*, 2013.

[6] Prakash S Raghavendra Poornalatha G. Web page prediction by clustering and integrated distance measure. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[7] A. Rajimol and G. Raju. Web access pattern mining, a survey. 2012.

[8] Xin Chen Yuhua Chen and Haoyi Chen. Improve on frequent access path algorithm in web page personalized recommendation model. *Internacional Conference on Information Science and Technology March 26-28,Nanjing, .Jiangsu, China*, 2011.

[9] Branko Markoski Zeljko Eremic, Dragica Radosav. Mining user access logs to optimize navigational structure of adaptive web sites. *11th IEEE International Symposium on Computational Intelligence and Informatics*, 2010.