



MODELO DE PREDICCIÓN SECUENCIAL DE WEB ACCESS LOG BASADO EN ALGORITMOS DE COMPRESIÓN Y MACHINE LEARNING

JAIME GUZMÁN

**Tesis para optar al título de Ingeniero Civil en Informática y
Telecomunicaciones**

Profesor guía: Adin Ramirez

**FACULTAD DE INGENIERÍA
ESCUELA DE INFORMÁTICA Y TELECOMUNICACIONES**

**Santiago, Chile
2015**



MODELO DE PREDICCIÓN SECUENCIAL DE WEB ACCESS LOG BASADO EN ALGORITMOS DE COMPRESIÓN Y MACHINE LEARNING

JAIME GUZMÁN

**Tesis para optar al título de Ingeniero Civil en Informática y
Telecomunicaciones**

**Profesor guía: Adin Ramirez
Comité: Francisco Claude
Comité: Darth Vader**

**FACULTAD DE INGENIERÍA
ESCUELA DE INFORMÁTICA Y TELECOMUNICACIONES**

**Santiago, Chile
2015**

📄 Jaime Guzmán
✉ mail@jguzman.cl

Utilice un par de oraciones para dedicar su tesis, o una frase de alguien importante.

Agradecimientos

Nota redactada sobriamente en la cual se agradece a quienes han colaborado en la elaboración del trabajo. No puede exceder más de una página.

Resumen

El siguiente documento comprende el estudio y creación de un algoritmo híbrido entre Machine Learning y Algoritmos de Loseless Data Compression. Dado que Internet crece cada día y los datos crecen en volúmenes del orden de los Terabytes, por lo cual es de interés usar técnicas de compresión para realizar un procesamiento de mayor información con la menor cantidad de recursos.

Hoy existe variadas tipos de web, redes sociales, microbloging, web informativas, etc. El contenido proporcionado a los usuarios finales ya no es estático y esto permite que los mismo puedan generar, aportar, modificar contenido, dado esto la ingeniería ofrece para construir web esta en constante evolución lo que ha ayudado a generar mas recursos para poder desarrollarla. Muchas de estas nuevas tecnologías han permitido entregar una mejor experiencia al momento de navegarla, aún cuando se ha creado un gran avance sobre el mismo esto no ha permitido crear Web que sean por sí mismas inteligente y puedan ir anticipando su comportamiento, para por ejemplo; disminuir la latencia desde que se abre una web ya visitada o desde que se navega dentro de un sitio con alta demanda; también desde el punto de vista de la arquitectura como servicio que las hospedan no se ha visto abarcada, dando un aspecto económico a los recursos utilizados. Si bien el crecimiento de los recursos de almacenamiento en la nube se encuentran un apogeo, las redes no crecen a la misma velocidad.

Este trabajo busca predecir el comportamiento de un usuario dentro de una web, usando técnicas de aprendizaje y algoritmo de compresión. Con este propósito se trabajará para crear un modelo híbrido.

Abstract

Abstract is the summary in english of the subject your are presenting in this thesis. Should not exceed one page.

Contenido

Resumen	i
Abstract	iii
Capítulo 1. Introducción	1
1.1. Contexto Preliminar	2
1.2. Definición del Problema	4
1.3. Algoritmos como servicio web	6
1.4. Predicción	7
1.4.1. Modelamiento de Eventos	7
1.4.2. Ventajas	8
1.5. Trabajos Relacionados	8
1.6. Descripción del Contenido y Contribuciones	10
Capítulo 2. Conceptos Básicos	11
2.0.1. Access Log	11
2.0.2. Árboles Trie	11
2.0.3. Alfabeto	11
2.0.4. Secuencias discretas	12
2.0.5. Lossless Data Compression	12
2.0.6. Motor de Predicción	12
2.0.7. Resilient Distributed Datasets	12
2.0.8. Data Source y Dataset	13
2.0.9. Cadenas de Markov	13
Capítulo 3. Predicciones sobre Web Access	15
3.1. Predictores de Estado finito	16
3.2. Modelos tradicionales	17
3.2.1. Limitaciones de los modelos tradicionales de Markov	17
Capítulo 4. Compresión y Machine Learning	19
4.1. Modelos de Compresión	20
4.1.1. Prediction by Partial Match (PPM)	21
4.1.2. Probabilistic Suffix Tree (PST)	21
4.1.3. Cadenas de Markov Dinámicas	21

4.1.4. Lempel & Ziv	22
Capítulo 5. Experimental	25
5.0.1. Datos Experimentales	26
5.0.2. Metodologías	27
5.0.3. Resultados Experimentales	27
5.1. Conclusiones y Resultados	28
Anexo A. Primer anexo	33
Anexo B. Segundo anexo	35

Lista de tablas

- 5.1. Para poder interpretar la entrada anterior se debe tener esta
relación de símbolos con respecto a las páginas 27

Lista de figuras

Capítulo 1

Introducción

Los nuevos avances tecnológicos y la inclusión de la ciencia de la computación en distintas campos han permitido hacer colecciones de datos muy grandes, las cuales se deben procesar, clasificar y analizar. Encontrar información útil dentro de estos grandes volúmenes de datos significa poder mejorar la decisiones a las que se pueden tomar dada a una base de conocimiento.

La minería de datos basado en web access se ha convertido en un área importante de investigación en los últimos años. Se puede utilizar para mejorar el rendimiento de la caché web, la detección de intereses de los usuarios y así recomendar páginas o bienes relacionados para los sitios web de comercio electrónico, mejorar los resultados de los motores de búsqueda y por ejemplo personalizar el contenido de la web con las preferencias deseadas para el usuarios . En la predicción de navegación web logrando entender el patrón de navegación del usuario y luego la predicción de las páginas siguientes es el principal problema que buscamos solucionar. Con un sistema de predicción fiable podemos ver la siguiente acción de los usuarios de Internet y tomar acciones.

Normalmente se ocupa muchos algoritmos de machine learning para poder hacer predicciones en variadas áreas. Hoy en día las aplicaciones en que los usuarios se enfrentan no pueden ser estáticas y sin tener un comportamiento que ayuden a predecir la forma en que actúan los usuarios. Esta área toma mucha relevancia al encontrarnos en una auge de la información generada por usuarios, redes sociales y variadas plataformas. Podemos mencionar

que esta es una de las razones para poder requerir disponer de herramientas de análisis y predicción que nos permitan saber como se comporta usuarios sobre una web, conocer la frecuencia en que se accede a un recurso en Internet, etc.

El momento en que un usuario entre a un web se establece una conexión cliente-servidor. Una gran cantidad de servicios proporcionan datos de accesos de los usuarios que acceden, sobre estos mismos se puede realizar variadas investigaciones sobre como predecir cual es la siguiente página que podrá visitar.

Las predicciones en los registros de web access han atraído una atención significativa en los últimos años. Muchas técnicas de recuperación de datos y algunos sistemas de personalización usan algoritmos de predicción. La mayoría de las aplicaciones actuales que predicen la siguiente página Web de un usuario tienen una componente en línea que hace la tarea de preparación de datos y una sección en línea que proporciona contenido personalizado a los usuarios en función de sus actividades de navegación actuales. En este trabajo se presenta un modelo de predicción en línea que se puede consumir como un servicio de API el cual da una integración ha variadas plataformas y sistemas que no tenga un componente en línea y con una buena precisión de la predicción. Nuestro algoritmo se basa en algoritmos LZ78 que están adaptados para modelar y representar la navegación secuencia del usuario en una web. Nuestro modelo disminuye la complejidad computacional, que es un problema grave en el desarrollo de sistemas de predicción en línea.

1.1. Contexto Preliminar

La Web crece constantemente y por ende su infraestructura, también la información que podemos obtener de los usuarios y concurrencia de los sistemas, la cual para usuarios finales se traduce en latencia y una mejor o peor experiencia de usuario. Paralelamente se suma un costo exponencial de recursos tanto en tecnologías de desarrollo como servicio que no son optimizados para poder dar una experiencia de usuario con calidad de servicio. Podemos reflexionar, entonces, que el no tener mayores recursos mejorará el rendimiento ni tampoco será lo óptimo para dar una calidad de servicio web, ya que el ancho de banda de Internet no crecerá a la misma proporción.

Adicionalmente, las tecnologías para la creación de web dinámicas e asíncronas han evolucionado a favor de traspasar la carga cliente. Hoy en día ya

se poseen lenguajes y framework que disminuyen considerablemente la carga de un servidor, por lo cual, un buen servicio web es proveer una balanceada carga dentro del cliente y el servidor, pero cuando se poseen un volumen de datos grandes es requerido tomar decisiones que los recursos y lenguajes no cubren, es ahí el interés de dar inteligencia a los servicios web.

Predecir los futuros accesos que un usuario tendrá en una determinada web. Entendiendo que la manera en que un usuario navega es su comportamiento registrado en una web, y que se puede analizar, estudiar y registrar mediante *Web Access Log* y a los cuales se puede hacer una minería de datos, Web Usage Mining. El por qué de hacer minería de datos es que cada día la web genera un innumerable cantidad de datos, por lo cual usar un algoritmo que se puedan operar comprimidos presenta un interés ya que además de disminuir el espacio físico o recursos utilizados, este se puede usar como un algoritmo de predicción y trabajar con una mayor cantidad de datos.

Los registros de accesos de manera procesada o pre-procesada, ayudaría a ingenieros de desarrollo web y diseñadores, como a usuarios finales a tener una experiencia de usuario mejor, disminuyendo por ejemplo la latencia en respuestas por parte de cada petición que realizan.

Hoy en día, las web no pueden ser simplemente dinámicas en contenidos, debe poseer una adaptabilidad a la demanda del usuario o proveer información que permita adaptarse a los eventos, por lo tanto, es de interés el profundizar en este tópico.

El interés en hacer un estudio sobre esto es poder hacer integraciones en áreas como compresión y la áreas que se dedican a completamente a hacer estudios sobre maquinas de aprendizaje. De por sí solas cada una se abordado independientemente lo cual es un interés converge en un problema en común que se puede resolver de manera eficiente.

Durante este trabajo se usarán técnicas de compresión de datos, se utilizará una infraestructura y patrón de implementación para modelos de Machine Learning. Adicionalmente toda la experimentación se llevará acabo disponibilizando los algoritmos y modelos como servicio REST, el cual se explicará mas adelante, ya mencionado lo anterior este trabajo es implementable en áreas productivas las cuales pueden presentar interés.

Definiremos que un usuarios es un la que se conecta a un servicio web, estos pueden ser paginas web informativas, redes sociales, etc. Este usuario establece una conexión directa con una pagina al momento de realizar esta operación, es posible almacenar datos los cuales vamos a llamar `.access log` ó

INTRODUCCIÓN

registros de accesos, durante el texto se mantendrán las referencias en inglés.

Un ejemplo de access log es el siguiente:

```
172.31.33.116 - - [26/Nov/2015:00:12:12 +0000] "HTTP/1.1" 200 1784 "http://localhost/home"
"Mozilla/5.0 (Linux; Android 5.1.1; SAMSUNG SM-G920I Build/LMY47X)
SamsungBrowser/3.2 Chrome/38.0.2125.102 Mobile Safari/537.36"
172.31.33.116 - - [26/Nov/2015:00:12:12 +0000] "HTTP/1.1" 200 179333 "http://localhost/news"
"Mozilla/5.0 (Linux; Android 5.1.1; SAMSUNG SM-G920I Build/LMY47X)
SamsungBrowser/3.2 Chrome/38.0.2125.102 Mobile Safari/537.36"
172.31.33.116 - - [26/Nov/2015:00:12:12 +0000] "HTTP/1.1" 200 24660 "http://localhost/health"
"
"Mozilla/5.0 (Linux; Android 5.1.1; SAMSUNG SM-G920I Build/LMY47X)
SamsungBrowser/3.2 Chrome/38.0.2125.102 Mobile Safari/537.36"
172.31.33.116 - - [26/Nov/2015:00:15:12 +0000] "HTTP/1.1" 200 24604 "http://localhost/sports"
"
"Mozilla/5.0 (Linux; Android 5.1.1; SAMSUNG SM-G920I Build/LMY47X)
SamsungBrowser/3.2 Chrome/38.0.2125.102 Mobile Safari/537.36"
172.31.33.116 - - [26/Nov/2015:00:20:12 +0000] "HTTP/1.1" 200 4860 "http://localhost/home"
"Mozilla/5.0 (Linux; Android 5.1.1; SAMSUNG SM-G920I Build/LMY47X)
SamsungBrowser/3.2 Chrome/38.0.2125.102 Mobile Safari/537.36"
172.31.33.116 - - [26/Nov/2015:00:22:19 +0000] "HTTP/1.1" 200 4841 "http://localhost/
finances"
"Mozilla/5.0 (Linux; Android 5.1.1; SAMSUNG SM-G920I Build/LMY47X)
SamsungBrowser/3.2 Chrome/38.0.2125.102 Mobile Safari/537.36"
```

El ejemplo anterior nos da mucha información interesante como la IP desde donde se conecta, el tipo de navegador, el dispositivo si es un telefono inteligente o un navegador de escritorio, la fecha en que se realizo el acceso y también lo mas relevante el destino del usuario.

1.2. Definición del Problema

El problema de la Predicción, ha surgido hace años y diversos investigadores han trabajado con distintos enfoques. Rissman y Langdom en los laboratorios Bell al realizar pruebas con un robot y hacer un experimento con un robot que tiraba una moneda compitiendo con humano, realizaba todos los calculos markovianos y calculos de las probabilidades condicionales para que cierto evento ocurra, a diferencia del sujeto que solo estaba esperando un resultado.

Predecir no es trivial, pero si podemos llegar a cercanos y minimizar el error de equivocarnos. Sin embargo, dos áreas han tratado de resolver el problema, LDC y Machine Learning de manera separada. Por parte de LDC los mayores problemas son que los predictores funcionan totalmente desconectados y no dan una de disponibilidad inmediata de los resultados, en cambios en el área de Machine Learning debemos crear un modelo para entrenar y luego poder generar una función predictiva.

Planteamos el problema de poder resolver tener un modelos híbrido juntando los patrones de cada área y disponerlo como un servicio inmediato

dando una predictibilidad inmediata que hoy en la industria es necesaria para poder hacer útiles estos algoritmo y dar un valor a los avances.

1.3. Algoritmos como servicio web

Los avances en el desarrollo de nuevas tecnología que brinden mejores experiencia en su uso del día a día, lleva a dar como podemos llevar varios escenarios idealizados a implementaciones empresariales reales. Es bastante común encontrar librerías que son bastante útiles para hacer Minería de Datos, Clustering y muchas operaciones que pueden recurrir en cálculos muy complejos pero no se pueden ofrecer como servicio. Ya en auge de las Infraestructuras Cloud, la capacidad de computo que se puede alcanzar no es un problema a lo que antes se enfrenta un Cientista de Datos.

Una API es un interfaz de programación de Aplicaciones que nos permiten intermediar el Servicio A con el Servicio B. Respectivamente A puede ser el proveedor y B el Demandante de servicios. Si quisiéramos analizar datos que se encuentran dentro de un servidor específico, estos se podrían consumir por esta interfaz. Existen variados clientes que nos permiten ayudar a esta comunicación, incluso se pueden utilizar por una terminal de Unix la cual es posible que mediante el programa *curl*, el cual no permita dialogar.

Ya se dispone de Infraestructura como servicio, Software como servicio, Plataformas como Servicios. Dado lo anterior ofrecer estos algoritmos para hacer que cada vez soluciones de desarrollo den valor agregado a la experiencia requerida por usuario. Es por lo cual hemos decidido utilizar una librería y framework que nos de esta posibilidad. Ofrecer algoritmos a la Industria como un servicio que ayuda de manera inteligente y multiplataforma que es el caso de una API.

Las ventajas de este patrón son heredados todo lo que ofrece un API, interoperabilidad, evitar problemas de Infraestructura, Resiliencia de Datos, Persistencia de Datos, Análisis y Procesamiento sin afectar un curso operacional de una aplicación.

Un ejemplo claro de esto son el análisis de datos en sistemas legados los cuales en plan de mejoras, no poseen la compatibilidad para poder realizarlo. Por otro lado no En sí los algormimos de Compresión ó algoritmos de Machine Learning tienden ocupar recursos los cuales pueden ser razones para no implementarlos.

1.4. Predicción

En esta sección se presenta formalmente la librería y framework Prediction.IO, es un servidor de Machine Learning Open Source para Cientistas de Datos y Desarrolladores que permite crear motores de predicción para ambientes de producción, con un bajo tiempo de entrenamiento y despliegue en ambientes productivos. Principalmente esta construido en Apache Spark, HBase y Spray.

Como ya se ha señalado este ambiente de trabajo se encuentra en un maduración completa que permite tanto disponer servidores con motores predictivos, como también toda una infraestructura distribuida para hacer que complejos algoritmos sean utilizados para solucionar problemas reales.

1.4.1. Modelamiento de Eventos

El modelamiento de eventos, es simplemente el hecho de poder llevar un feature que es del mundo del ML, es en fin una representación de como se debe tener la data de manera RDD para poder acceder a ella posteriormente

Un evento lo definiremos como entidad que nos permite dar una representación temporalizada de información la cual será procesada por un motor de predicción. Analizaremos los eventos que un usuarios realiza para poder acceder a una web. Adicionalmente cuando cada usuario ingresa a una web automáticamente este genera un sesión, la cual es desde que llega hasta que abandona la web.

Como ya se ha mencionado en sección anteriores esta información esta totalmente depurada y entregada por los access log, los cuales a efectos de temporalidad nos interesa conocer la secuencialidad discreta de estos accesos.

El modelamiento que se realizará contempla que el usuario :

- Tipo de Evento: Visitar
- Entidad que ejecuta el evento: Usuario
- Propiedades:
 1. Pagina actual
 2. Pagina siguiente
 3. Cierre de Sesión

El interés de tener un modelo totalmente atómico es poder contemplar la información que nos entrega, destacando sus variables y propiedades como restricciones.

- Persistencia en el modelo

1.4.2. Ventajas

El crecimiento de este tipo de arquitectura es tanto horizontal como vertical

Mix in and implement this trait if your model cannot be persisted by PredictionIO automatically. A companion object extending IPersistentModelLoader is required for PredictionIO to load the persisted model automatically during deployment.

Notice that models generated by PAlgorithm cannot be persisted automatically by nature and must implement these traits if model persistence is desired.

Comentar sobre DASE Data Source Algoritim Serving Engine

A Resilient Distributed Dataset (RDD), the basic abstraction in Spark. Represents an immutable, partitioned collection of elements that can be operated on in parallel. This class contains the basic operations available on all RDDs, such as map, filter, and persist. In addition, org.apache.spark.rdd.PairRDDFunctions contains operations available only on RDDs of key-value pairs, such as groupByKey and join; org.apache.spark.rdd.DoubleRDDFunctions contains operations available only on RDDs of Doubles; and org.apache.spark.rdd.SequenceFileRDDFunctions contains operations available on RDDs that can be saved as SequenceFiles. All operations are automatically available on any RDD of the right type (e.g. RDD[(Int, Int)] through implicit.

Internally, each RDD is characterized by five main properties:

A list of partitions A function for computing each split A list of dependencies on other RDDs Optionally, a Partitioner for key-value RDDs (e.g. to say that the RDD is hash-partitioned) Optionally, a list of preferred locations to compute each split on (e.g. block locations for an HDFS file) All of the scheduling and execution in Spark is done based on these methods, allowing each RDD to implement its own way of com

1.5. Trabajos Relacionados

En la literatura, el tema de la predicción en la web se ha presentado como un tema concurrente, y ha sido abarcado por varios autores. Tenemos los

siguientes trabajos de interés:

1. Dongshan y Junyi [1] 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 una representación de *Tree Order Model*, el cual es un híbrido entre un modelo de markov tradicional y una representación de árbol, bautizada como HTMM (por sus siglas en inglés, *Hybrid-Order Tree Markov Model*). Su modelo fue presentado en 2002, y da una 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. Domenech *et al.* [2], muestran un estudio de los rendimientos de técnicas de recuperación de datos. Las mismas se pueden utilizar para dar una entrada ideal a algoritmos de aprendizaje o algoritmos de predicción. Los conceptos más importantes son las nuevas variables de caracterización, temporalidad, espacio y geografía, que se le suman a la predicción. Además de comenzar un trabajo más elaborado de como tomar una predicción, se introducen conceptos como predicciones genéricas o específicas, variables de uso de recursos a nivel de red ó nivel procesamiento. Finalmente, se presenta un modelo predictivo que 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. Chen *et al.* [3] dan una nueva perspectiva enfocada a entregar una clara recomendación a los usuarios basada en la misma propuesta de este proyecto, los access log. El primer análisis realizado por los autores cubre las reglas asociativas que requiere un sistema de recomendación, pero en las pruebas propiamente tales encuentran que el análisis de los patrones detectados dan una representación clara de como optimizar la web, y finalmente mediante sus pruebas logran una recomendación de calidad.
4. Rajimol y Raju [4] minaron los patrones de los accesos web, donde el enfoque es usar los registros de acceso para crear subsecuencias y realizar comparaciones. La literatura presenta un interés para poder anticipar el patrón de comportamiento de la web.
5. Kewen [5] realizó un análisis más profundo del *web usage minning*.

Parte de la importancia de este trabajo, es que después de minar los registros de accesos, logran reducir la “*bad data*”.

6. Poornalatha y Raghavendra [6] establecen que se pueden utilizar máquinas de aprendizaje para predecir basándose en distintas entre clusters. Estos autores, al igual que Domenech *et al.* [2] y Dongshan y Junyi [1], comparan el objetivo de optimizar los recursos tanto en redes (disminución de latencia) y experiencia de usuario.
7. Claude *et al.* [7] presentan una estructura de representación eficiente que permite dar una representación de *web access log* y ofrecen las operaciones básicas de WUM.

1.6. Descripción del Contenido y Contribuciones

Capítulo 2

Conceptos Básicos

En este capítulo se introducen los conceptos principales que se trabajarán en los siguientes capítulos:

2.0.1. Access Log

Son los registros que se almacenan en un servidor, los cuales dependiendo del sistema operativo pueden tener mayor o menor información. Cuando los usuarios acceden a los sitios web, estos suelen dejar una gran cantidad de información de acceso, la cual si es extraída razonable puede ayudar a los administradores de sitio web para obtener acceso a los patrones de los usuarios.

2.0.2. Árboles Trie

Son estructuras de datos de tipo de árbol que almacenan datos en nodos y es de muy fácil la recuperación de información de estos mismo. Sus características generales es ser un conjunto de llaves las cuales se representan en el árbol y sus nodos internos representan la información, en nuestro caso una caracter o string de tamaño 1.

2.0.3. Alfabeto

Un alfabeto discreto A consiste en $M > 1$ símbolos. Dado un volumen de datos experimental, nuestro alfabeto es representado simbólicamente como

la representación de un nodo de contenido de un sitio web. Donde A , puede ser definido como la página inicial. Este alfabeto es finito y acotado por la minería de datos de uso web.

2.0.4. Secuencias discretas

Definimos una secuencia de accesos discreta y finita, dado los accesos que tiene un usuario frente a una web, lo anterior es acotado por el concepto de sesión, el cual es desde que se inicia la navegación, es decir secuencia de tamaño $Seq \leq 1$ y de tamaño no superior a un alfabeto A .

2.0.5. Lossless Data Compression

La compresión sin pérdida o LDC, es el arte de poder comprimir bits y poder hacer el proceso inverso, es decir tener la posibilidad de codificar y decodificar. En capítulos posteriores introduciremos más el tema compresión y como este ayuda a crear un modelo de predicción.

2.0.6. Motor de Predicción

Es la parte fundamental de un sistema dirigida a adivinar el futuro acceso de un usuario. La salida del motor de predicción es la siguiente página, que se compone de un símbolo representando la dirección URL o una sección en particular de un cierta web.

2.0.7. Resilient Distributed Datasets

Los RDD, permiten que en un servidor de Machine Learning sea posible mantener un modelo o motor de aprendizaje persistente sin importar el flujo en que se encuentre.

Esta estructura es fundamental dentro de la librería que se introducirá más adelante, Apache Spark. Esta estructura es una colección distribuida de objetos Inmutable, cada set de datos en un RDD es dividido en particiones lógicas, las cuales pueden ser computadas en distintos Clusters. Los RDD pueden contener cualquier tipo de objeto de los siguientes lenguajes: Python, Scala y Java, incluyendo clases definidas por el usuario.

Formalmente los RDD son solo de lectura, una colección de objetos particionada. Estos pueden ser creados a través de operaciones determinísticas en una cierta tabla o un almacenamiento externo o otra RDD. Otra de las

características de los RDD, es que son colecciones de elementos tolerantes a fallas que puede ser operadas en si mismas o en paralelo. Apache Spark hace el uso del concepto de RDD para lograr rapidez y eficiencia en las operaciones de MapReduce, de ser requeridas. Destacamos la escalabilidad de esta librería para un gran nivel de computo, pero en este trabajo no se explicará el uso de Spark, pero si se utilizarán algunos conceptos como RDD y otros.

2.0.8. Data Source y Dataset

Ambos conceptos están focalizados en proveer información tanto para el servidor de Machine Learning, como para el procesamiento y analisis. En este trabajo los dataset a estudiar son un los registros de accesos de la web española Prisa, los cuales representan 1.000.000 de registros correspondientes.

Nuestro set de datos esta basado en los registros (webaccess log), ya previamente depurados con una representación numérica desde 0 hasta 17

frontpage news tech local opinion on-air misc weather msn-news health living business msn-sports sports summary bbs travel

2.0.9. Cadenas de Markov

Capítulo 3

Predicciones sobre Web Access

Dado los nuevos sistemas es cada vez más común conocer sistemas Inteligentes y diversos en variadas áreas, una de estas cualidades de los nuevos sistemas este tener la posibilidad de predecir ocurrencia de eventos para poder adaptarse y tener versatilidad al tomar decisiones en variadas situaciones. Aún mas necesaria es esta propiedad en problemas que requieren predicciones secuenciales, es decir dada una secuencia de eventos, como poder predecir el siguiente evento basado en nuestro conocimiento histórico limitado.

Para cualquier secuencia de eventos, estas se pueden modelar como proceso estocásticos, esto algoritmos emplean Modelos de Markov para optimizar las predicciones del siguiente simbolo, en cualquier secuencia estocástica. Otros escenarios requieren que algoritmo de predicción sea capaz de incrementar su recuperación de información y poder dar resultados de forma inmediata, es decir predicciones *online* .

El problema de la predicción secuencial se puede establecer de la siguiente manera, dado una secuencia de simbolos x_1, x_2, \dots, x_n , ¿Cuál es el siguiente simbolo x_{i+1} ? Señalado el problema la literatura converge a que gracias Rissanen y Langdon un buen compresor, o mejor dicho que cualquier compresor de datos sin pérdida se aproxima a un *predictor*.

Para poder crear un predictor acorde a la Teoría de la Información, un predictor que construye un modelo cuya entropía aproxima a la de la fuente de datos, consigue una mayor precisión predictiva.

Las predicciones son un area importante dentro del dominio de las Machi-

ne Learning y la Inteligencia Artificial, las cuales pueden ofrecer un sistema de inteligencia que las aplicaciones necesitan para un optimio desempeño, también ayudan a dar información para la elección de decisiones. Ciertos dominios requieren que la predicción se pueda realizar en las secuencias de eventos que por lo general se puede modelar como un proceso estocástico.

Nuestro interés se focaliza en las predicciones de secuencias discretas, en el cual queremos demostrar la convergencia en el cual un modelo de compresión, como el caso de LZ78 que se explicará mas detalladamente en el capitulo posterior. La eficiencia de un algoritmo de compresión ofrece una nueva perspectiva a las predicciones, la cual ha despertado un interés en investigadores.

Nos enfocaremos en el caso del los acceso que un usuario realiza a un sitio web, el tiempo en el que pasa en este es registrado por el servidor, en esta investigación no se requiere indagar en temas de Information Retrieval, ya que se entrega una colección de datos ya procesada, la cual es representa secuencias de acceso por parte de usuarios.

Nuestro interés se presetna en dada un sequencia acceso, un usuario entra accede a la web, dado sus accesos web cual será la predicción de su sesión. Dentro de los registros existen respaldo de sesiones de usuarios, páginas no encontradas, accesos denegados y otra información en relación al funcionamiento. Sin importar el tipo de servidor que utilicemos podemos identificar a usuarios con distintas técnicas y/o combinaciones, por ejemplo podemos interpretar que el *request* que ha realizado el usuario mas su *session* nos da la información necesaria para poder crear un modelo predictivo.

Dado lo anterior podemos hacer minería de los datos que un servidor ha recolectado durante un periodo de tiempo, de lo anterior podemos mencionar que existen varias técnicas para poder hacer *Web Access Pattern* y también "Web Usage Mining", de sus siglas en ingles es minería del uso de una *web*. Nuestro interés no es abordar este tema pero si explicarlo para poder realizar un estudio predictivo de secuencias de acceso, con el objetivo de poder predecir el siguiente comportamiento que tiene un usuario al momento de visitar un sitio web.

3.1. Predictores de Estado finito

Si se tiene un secuencia estocástica, $x_n = x, x, \dots, x.$, en un tiempo t el predictor debe saber cual es el siguiente simbolo, basdo en su historia o la

frecuencia de ocurrencia que se vaya entrenando.

Si llamamos al siguiente simbolo b_t , diremos que el resultado de nuestro predictor es entregar este valor. Dado esto, existe una función de perdida asociada $L(b_t, x_t)$ para cada predicción realizada.

El objetivo de cada predictor es tener una función de minimización tal que minimice la fracción de predicciones erróneas, a lo anterior lo llamaremos T que será:

3.2. Modelos tradicionales

Primero asumimos que cada evento por si

En secciones anteriores se ha hablado de como los eventos secuenciales se pueden modelar como un proceso estocástico.

Ciertamente el uso de reconocer frecuencia patrones secuenciales, reglas de asociación puede requerir mucha información y entregar mejor precisión en la información de los usuarios y su comportamiento en la web. Pero un enfoque en que la temporalidad de los datos se debe modelar con un modelo de Markov es la clave para poder realizar esto, y estos ya han ayudado a predecir los accesos de los usuarios, como señala pero en la práctica existen muchas limitaciones técnicas que permiten que se implementen.

3.2.1. Limitaciones de los modelos tradicionales de Markov

Los modelos tradicionales de Markov predicen la siguiente página Web que un usuario puede acceder considerando el acceso más probable, iterando coincidir su secuencia de acceso actual con secuencias de acceso Web histórica.

Usando estos modelos se ha comparado los investigadores comparan el máximo de elementos prefijos de cada secuencia histórica de web access con los elementos sufijos de la misma longitud de secuencia de web access actual del usuario y obteniendo secuencias históricas con la probabilidad más alta de elementos que coinciden.

El modelo de Markov de orden cero es la tasa base de probabilidad incondicional, la cual es la probabilidad de la página visitada, dada por

$$p(x_n) = Pr(X_n), \quad (3.1)$$

donde x_n es x_n y X_n es otra X_n .

El modelo de Markov de orden uno observa la probabilidad de la transición de un pagina a otra, la podemos interpretar así:

$$p(x_2|x_1) = Pr(X_2 = x_2|X_1 = x_1) \quad (3.2)$$

El K-ésimo orden del Modelo de Markov considera la probabilidad condicional que un usuario cambie a una nueva n-ésima página dado su anterior pagina visitada, teniendo que $k = n - 1$ paginas vistas:

$$p(x_n|x_{n-1}, \dots, x_{n-k}) = Pr(X_n = x_n|X_{n-1} = x_{n-1}, \dots, X_{n-k} = x_{n-k}) \quad (3.3)$$

Modelos de Markov, en (3.3), de orden inferior no pueden predecir con éxito total el futuro de los web access log, ya que no se ven lo suficientemente atrás en el pasado para discriminar el modo en que se comporta el usuario. Este comportamiento de los usuarios tanto requiere de buenas predicciones los cuales a su vez requieren modelos de Markov de orden superior, pero los modelos de orden superior resultaran de alta complejidad en espacio de estado y cobertura.

Modelos con mayor Orden de estados son distintas combinaciones de las acciones observadas en un set de datos, entonces el numero de estados tiende a crecer exponencial-mente al igual que el Orden del modelo.

Este aumento puede limitar significativamente la aplicabilidad de los modelos de Markov para aplicaciones en las que las predicciones rápidas son críticas para el rendimiento en tiempo real o para aplicaciones con restricciones de uso de memoria. Además, muchos ejemplos en los test podrían no tener estados correspondientes en los modelos de Markov de mayor orden, por lo que reduciría su alcance.

Capítulo 4

Compresión y Machine Learning

El uso de Algoritmos de compresión en tareas de Machine Learning como clusterización y clasificación ha tenido presencia en variados campos, con la intención de reducir problemas de selección explícita de ciertas características que se usan en estudios y algoritmos de Machine Learning.

Un punto de vista de esta inclusión de áreas muestra que los algoritmos de compresión mapean implícitamente string en representaciones vectoriales de dichas características las cuales son cotas superiores. Podemos señalar que los trabajos como Langdon y Rissman han sido claves para determinar que modelos de compresión pueden usarse para realizar predicciones y esto ha sido una área de gran interés en Machine Learning.

Un punto de vista alternativo muestra algoritmos de compresión mapean implícitamente cuerdas en vectores implícitos espacio de características, y por compresión similitud medidas basadas computan similitud dentro de estos espacios de características.

Esta idea de usar algoritmos de compresión en máquinas de aprendizaje no es nueva, pero no ha sido explotada mayormente explorada.

Los algoritmos de compresión han sido estudiados e investigados por durante varios años, la motivación fundamental es poder optimizar el espacio, para mayor uso o almacenamiento de datos. Estos algoritmos se encuentran sin saberlo en nuestro día a día, desde el núcleo de un sistema operativo como linux hasta por ejemplo los formatos *zip*, *rar*, *7z*, también en formatos de imágenes y audios, etc. los cuales son útiles para poder optimizar una trans-

ferencia de archivos de un equipo a otro mediante Internet o simplemente comprimir datos para respaldar en dispositivos físicos, etc.

La motivación de profundizar en el área de compresión de datos se debe a una de la razones mencionadas con anterioridad, Internet. Esta red de redes, constante crea nuevos contenidos, registros, imágenes etc. los cuales no es conveniente mover de un lugar a otro mediante un transferencia directa, estos archivos crecen innumerablemente y aquí es uno de los mayores aporte que poseen los algoritmos de compresión con relación a nuestra red de redes.

A diferencia del volúmenes de datos las infraestructura de redes y su velocidad no crece directamente proporcional, que esto genera un sin fin de problemas para los usuarios e industria web. La latencia es el tiempo de respuesta que demora un usuario en solicitar, hacer un *REQUEST*, a un servidor, simplemente es un el tiempo de respuesta desde iniciada una acción demandada. Uno de los grandes ejemplos que tenemos en la web es la proliferación de archivos comprimidos para su descarga, los cuales en su interior poseen variados recursos multimedia, texto, etc.

Las propiedades de estos algoritmo no solo permiten juntar un set de archivos y lograr un tasa de compresión optima para ser transmitido por Internet, también pueden ayudar a realizar análisis en grandes volúmenes de información, por ejemplo; el análisis de texto, clasificación de proteínas, moderación de contenidos en web y predicciones del comportamiento de usuarios que navegan en un sitio de Internet. Sobre este último punto es nuestro mayor interés debido a ya las antes mencionadas similitudes que poseen un algoritmo de compresión y un modelo variable de Markov.

Para introducir el camino se debe presentar formalmente los algoritmos de compresión y su clasificación más general. Entre ellos tenemos los algoritmos con pérdida y sin pérdida, nos enfocaremos en los algoritmos *Lossless Compression Algorithm*, algoritmos de compresión sin pérdida.

Learning of sequential data continues to be a fundamental task and a challenge in pattern recognition and machine learning.

4.1. Modelos de Compresión

Existen muchos modelos y algoritmos de compresión, nuestro enfoque es usar los algoritmos de compresión que tengan un espacio vectorial de características conjunto con Machine Learning y además tengan propiedades para ser candidatos a un predictor.

4.1.1. Prediction by Partial Match (PPM)

El algoritmo de predicción por certeza parcial es considerado uno de los mejores algoritmos del tipo *Lossless Compression Algorithms*. El algoritmo requiere un tope superior D en el máximo del orden de Markov de un modelo variable de Markov (*VMM*) para construirse. PPM maneja el problema de frecuencia cero usando dos mecanismos llamados

- Escape
- Exclusion

For a method that considers different orders of models, we turn once again to data compression and the Prediction by Partial Match (PPM) family of predictors. This has been used to great effect in (Ref. 4), for a predictive framework based on LZ78. PPM algorithms consider different-order Markov models in order to build a probability distribution by weighting different-order models appropriately. In our predictive scenario, Active LeZi builds an order- k Markov model. We now employ the PPM strategy of exclusion (Ref. 7) to gather information from models of order 1 through k to assign the next symbol its probability value. This method is illustrated by considering the example sequence used in the previous sections: “aaababbbbbaabccddc-baaaa”. The window maintained by Active LeZi represents the set of contexts used to compute the probability of the next symbol. In our example, the last phrase “aaa” (which is also the current ALZ window) is used. Within this phrase, the contexts that can be used are all suffixes within the phrase, except the window itself (i.e. “aa”, “a”, and the null context).

4.1.2. Probabilistic Suffix Tree (PST)

Los árboles de sufijos implementados como un algoritmo de predicción intentan construir el único y mejor VMM con límite superior D , acorde a la secuencia de entrenamiento de entrada. Esto asume que un límite superior a la orden de Markov de un "fuente certera"^{es} conocida como *learner*.

4.1.3. Cadenas de Markov Dinámicas

Los Algoritmos DMC ó *Dinamyc Markov Compression* son modelos de información con máquinas de estados finitos. Las asociaciones están hechas entre todos los símbolos posibles en el alfabeto origen y la distribución de

probabilidad sobre todos los simbolos en el alfabeto. Esta distribución de probabilidad es usada para predecir el siguiente digito binario. Los *DMC* comienzan en un estado ya previamente definido, cambiando de estado cuando nuevos bits son leídos desde la entrada. La frecuencia de transmisión ya sea un 0 or un 1 son sumados cuando un nuevo simbolo entra. La estructura puede también ser actualizada usando *state cloning method*.

4.1.4. Lempel & Ziv

LZ78 algorithm is proposed by Jacob Ziv and Abraham Lempel in 1977 [15]. An online prediction method needs not rely on time-consuming preprocessing of the available historical data in order to build a prediction model. The preprocessing is done when we have a new request. LZW and LZ78 basically are lossless data compression algorithms with good functionality. The most important part of these algorithms is the dictionary construction algorithm that we use it for creating the prediction model.

```

initialize dictionary := null
initialize phrase w := null
loop
wait for next symbol v
if ((w.v) in dictionary):
w := w.v
else
add (w.v) to dictionary
w := null
increment frequency for every possible prefix of phrase

endif
forever

```

LZ78 is a lossless compression algorithm. Fig.2 shows that how the dictionary constructed from sequences using LZ78. In web environment we use user web page requests sequence as input sequence of LZ78 algorithm. Fig.3 shows how prediction tree is constructed. In Fig.2 and 3 variable w is sequence that is saved in each user session. This algorithm can insert sequences with long length, but generally total number of sequences that inserted in

tree is less than PPM algorithm. We explain this algorithm with an example. Suppose the user requests the pages ABABCBC sequentially. If we use the LZ78 algorithm, then the A, B, AB, C and BC should be inserted in the tree. In Table 1 the first row shows the user requests. The second row shows the sequences inserted in the tree and the third row shows the sequences that maintained in active user session. When a sequence is inserted in the tree the weights of edges that represent the pass from the root to the last request of sequence is incremented. Now assume that user B requests the sequence of pages ABCABCD. Table 2 shows the results. If user A requests ABABCBC

Problems in LZ78 parsing Any practical implementation of LZ78 suffers from the following two drawbacks: i. In any LZ parsing of an input string, all the information crossing phrase boundaries is lost. In many situations, there will be significant patterns crossing phrase boundaries, and these patterns will affect the next symbol in the sequence. ii. The convergence rate of LZ78 to the optimal predictability as defined above is slow. The results outlined in (Ref. 1) by Feder, et al state that LZ78 asymptotically approaches optimal predictability.

Capítulo 5

Experimental

Lo que se busca es usar como un modelo de predicción un algoritmo de la familia de compresores de Lempel Ziv. Se usará para predecir secuencias finitas discretas de accesos web. Al momento de generar el árbol este creará una representación trie de un diccionario de símbolos, el cual no utilizaremos para poder hacer predicciones.

En base a lo anterior y teniendo funcionando el algoritmo para crear un modelo de predicción, la integraremos con el servidor de Machine Learning, Prediction.IO que ya se ha explicado anteriormente.

Para los ambientes experimentales se han dispuestos dos máquinas para realizar las pruebas.

Máquinas

- Procesador 2,8 GHz Intel Core i7, 16 GB de Memoria RAM y Sistema Operativo OSX
- Procesadores Intel Xeon E5-2670 v2 (Ivy Bridge) de alta frecuencia 32 vCPU, 244 GB de Memoria RAM y Sistema Operativo Ubuntu 14.14

Software

- C++11
- Java 1.8

- Java(TM) SE Runtime Environment (build 1.8)
- Java HotSpot(TM) 64-Bit Server VM (build 25.51 – b03, mixed mode)
- Scala code runner version 2.11.7 – Copyright 2002-2013, LAMP/EPFL
- SBT 0.13.9
- Python 2.7.10
- Prediction.IO 0.9.4
- Elasticsearch 1.4.4
- Apache Spark-1.4.1
- Hbase 1.0.0
- Zookeeper

5.0.1. Datos Experimentales

Se usará las secuencias disponibles de acceso web pública de los sitios Msnbc. El set de datos provienen de los registros de un servidor IIS (Internet Information Services) msnbc.com de un día completo de la fecha 28 de Septiembre de 1999. Contiene secuencias de acceso web de 989,818 usuarios con un promedio de 5,7 categorías Página web visitas por secuencia, el tamaño de letras de este conjunto de datos es $\sigma = 17$.

Ejemplo de estructura de Dataset MSNBC :

```
% Different categories found in input file:

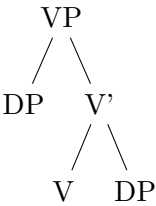
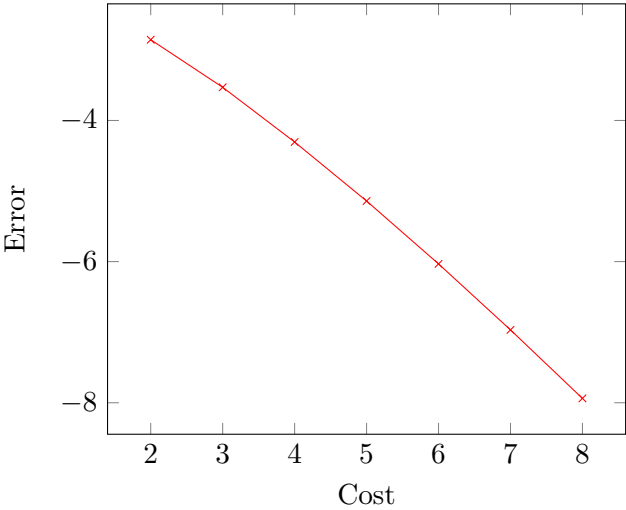
frontpage news tech local opinion on-air misc weather msn-news health living business msn-
sports sports summary bbs travel

% Sequences:

A A
B
C B B D B B B C C
E
A
F
A A
F
F G G G F F H H H H
F I D D D J C J E J D D D
A A A K A A A
L L
A A
H H H H H H
```

Tabla 5.1. Para poder interpretar la entrada anterior se debe tener esta relación de símbolos con respecto a las páginas

Page	Symbol
frontpage	A
news	B
tech	C
local	D
opinion	E
on-air	F
misc	G
weather	H
msn-news	I
health	J
living	K
business	L
msn-sports	M
sports	N
summary	O
bbs	P
travel	Q



5.0.2. Metodologías

5.0.3. Resultados Experimentales

1. Experimento con Largo de Ventana

Cual podria ser el optimo de la ventana Que ventana o cual corre mejor

2. Experimento con el Tamaño del Set de datos de Entrenamiento

3. Experimento con Predicciones con Ruido

4. Cross Validation

5. Accuracy

6. Experimento comparativo con Frequent Sequence Pattern

7. Decision Tree

8. Association Rules

5.1. Conclusiones y Resultados

Referencias bibliográficas

- [1] X. Dongshan and S. Junyi, “A new markov model for web access prediction,” *Computing in Science Engineering*, vol. 4, pp. 34–39, Nov 2002.
- [2] J. Domènech, J. A. Gil, J. Sahuquillo, and A. Pont, “Web prefetching performance metrics: A survey,” *Performance Evaluation*, vol. 63, no. 9–10, pp. 988–1004, 2006.
- [3] Y. Chen, X. Chen, and H. Chen, “Improve on frequent access path algorithm in web page personalized recommendation model,” pp. 83–86, March 2011.
- [4] A. Rajimol and G. Raju, “Web access pattern mining, a survey,” 2012.
- [5] L. Kewen, “Analysis of preprocessing methods for web usage data,” *International Conference on Measurement, Information and Control (MIC)*, 2012.
- [6] G. Poornalatha and P. Raghavendra, “Web page prediction by clustering and integrated distance measure,” in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pp. 1349–1354, Aug 2012.
- [7] F. Claude, R. Konow, and G. Navarro, “Efficient indexing and representation of web access logs,” 2014.
- [8] R. Begleiter, R. El-Yaniv, and G. Yona, “On prediction using variable order markov models,” *J. Artif. Int. Res.*, vol. 22, pp. 385–421, Dec. 2004.
- [9] Z. Eremic, D. Radosav, and B. Markoski, “Mining user access logs to optimize navigational structure of adaptive web sites,” pp. 271–276, Nov 2010.
- [10] K. Gopalratnam and D. Cook, “Online sequential prediction via incremental parsing: The active lezi algorithm,” *Intelligent Systems, IEEE*, vol. 22, pp. 52–58, Jan 2007.
- [11] T. Gueniche, P. Fournier-Viger, R. Raman, and V. Tseng, “Cpt+: Decreasing the time/space complexity of the compact prediction tree,” in

REFERENCIAS BIBLIOGRÁFICAS

- Advances in Knowledge Discovery and Data Mining* (T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, D. Cheung, and H. Motoda, eds.), vol. 9078 of *Lecture Notes in Computer Science*, pp. 625–636, Springer International Publishing, 2015.
- [12] B. Kotiyal, A. Kumar, B. Pant, R. Goudar, S. Chauhan, and S. Juneja, “User behavior analysis in web log through comparative study of eclat and apriori,” pp. 421–426, Jan 2013.
- [13] J. Li, “Research of analysis of user behavior based on web log,” pp. 601–604, June 2013.
- [14] A. Moghaddam and E. Kabir, “Dynamic and memory efficient web page prediction model using lz78 and lzw algorithms,” pp. 676–681, Oct 2009.
- [15] J. de-la Pena-Sordo, I. Pastor-Lopez, I. Santos, and P. G. Bringas, “Using compression models for filtering troll comments,” pp. 655–660, June 2015.
- [16] D. Sculley and C. Brodley, “Compression and machine learning: a new perspective on feature space vectors,” pp. 332–341, March 2006.
- [17] X. Shan and H. Sun, “The research of web users’ behavior mining based on association rules,” pp. 7415–7418, Aug 2011.

ANEXOS

Anexo A

Primer anexo

Anexo B

Segundo anexo
