

NOMBRE CÓDIGO - CATEGORIZACIÓN

Kiara Lucas Herrera¹

¹Carrera de Ingeniería en Sistemas Computacionales
Universidad de Guayaquil



Figura 1. Logo de Amazon

I. ABSTRACTO

El siguiente artículo habla acerca de la jerarquización de productos existentes en la Página de Amazon basado en el estudio de Data Mining; nos encargaremos de descargar todos los datos que existen en la página y por diferentes métodos estudiados ir procesando toda la información obtenida para llegar a agrupar cada artículo al conjunto que pertenece según las características específicas. ¿Cómo lo haremos?. A continuación explicaremos brevemente los métodos a utilizar:

METODO 1- Obtención de los datos; como ya se habían descargado anteriormente los datos de Amazon lo que hicimos fue obtener esa Información y analizarla para comenzar a resolver la problemática.

METODO 2- Utilizando MapReduce lograremos optimizar la información y sacar los datos necesarios para resolver nuestro proyecto desechando información que no formará parte del algoritmo a utilizar.

METODO 3- Resolución del algoritmo. Como resultado final queremos obtener varios cluster de forma jerárquica; para ello utilizaremos un algoritmo específico que se encarga de darnos el resultado que estamos buscando.

METODO 4- Agrupación de clusters. Al haber utilizado el algoritmo jerárquico tendremos los

diferentes cluster de productos de la página de Amazon, llegando al resultado propuesto.

II. INTRODUCCIÓN

Amazon es una empresa dedicada a la venta de productos por internet, su sitio web a diario es sumamente visitada por cientos de personas que quieren adquirir un producto en especial; pero muchas veces se complica hallar el indicado debido a que no están totalmente jerarquizados para poder elegir la mejor opción entre los productos que necesitamos.

Pensando en aquella problemática hemos propuesto realizar la **Clasificación Jerárquica de los productos existentes en Amazon**, para que aquellos clientes tengan la posibilidad de elegir la categoría que necesitan. Al poder obtener la categoría podrán buscar en ella el producto deseado según las características que éste posea optimizando la información que no es necesaria y dejando los datos precisos para que el cliente quede a gusto con lo que necesita comprar. De esta manera se reducen ciertos datos que se añaden de más y que sólo retardan la búsqueda del producto ya que son irrelevantes para el cliente. Amazon es una de las páginas más cotizadas en ventas por internet a nivel mundial por esta razón la elegimos para realizar el estudio de Data Mining¹.

III. TRABAJOS RELACIONADOS

Hemos tomado como ayuda o referencia un estudio de Algoritmo de Clustering basado en entropía para descubrir grupos en atributos de

tipo mixto; el cual nos explica detalladamente cómo utilizar el algoritmo que nosotros vamos a plantear para resolver la problemática expuesta al principio, además nos enseña los tipos de datos que pueden existir en un Clustering despejando así todas las dudas que podamos haber tenido acerca de los detalles que utilizaremos en el proceso de resolución de jerarquización en Amazon. Este estudio nos muestra cómo podemos resolver el algoritmo que nos dará los resultados que esperamos, enseñándonos con gráficos y procedimientos paso a paso el uso del algoritmo jerárquico que necesitamos entender para nuestro proyecto. Hemos visto en el documento de referencia que estamos utilizando la metodología que utiliza el algoritmo jerárquico para resolver los diferentes clustering que nos dará como resultado final al momento de implementarlo.

(tesisEdnaHernandez.pdf)

IV. DATOS

Los datos serán tomados de un dataset que descargamos desde el sitio web de la Universidad de Stanford.

Conjunto de datos (dataset)

Es la materia prima del sistema de predicción. Es el histórico de datos que se usa para entrenar al sistema que detecta los patrones. El conjunto de datos se compone de instancias, y las instancias de factores, características o propiedades.

Representación de los Datos Los datos son una colección de entidades mapeadas en un dominio de interés. Su representación simbólica se basa en las relaciones existentes entre un conjunto de atributos que describen a un conjunto de objetos. Los atributos representan las propiedades y características básicas de los objetos. Son también conocidos como: variables, campos o características 3, 4, 5, 7 .

Tipos de DataSets: En los últimos años, investigadores han estudiado distintos métodos para un efectivo y eficiente análisis de clustering en datasets con distintos tipos de datos. En función del tipo de dato contenido en el dataset, se han presentado 3 categorías generales:

1. **Numéricos:** Pueden ser analizados en función de las características inherentemente geométricas de los datos. Comúnmente se

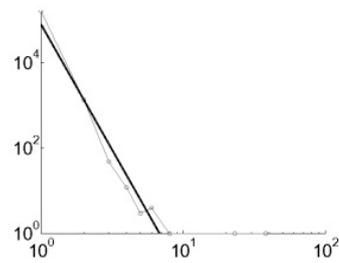


Figura 2. VIDEO

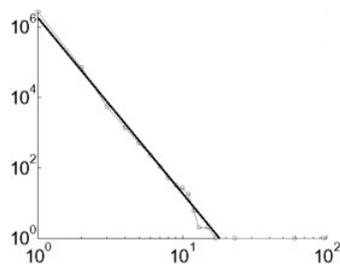


Figura 3. LIBRO

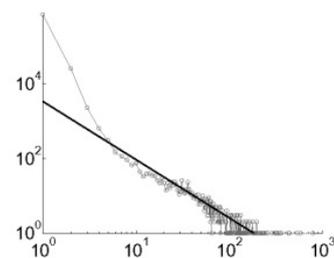


Figura 4. DVD

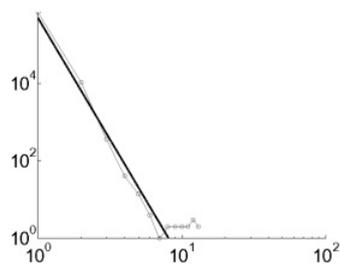


Figura 5. MUSIC

utilizan medidas geométricas (ej. funciones de distancias geométricas).

1. **Categoricos:** Se analizan de acuerdo a las características cualitativas de los datos. Se utilizan medidas de similitud y análisis de frecuencia para evaluar la estructura representativa de los datos.
1. **Mixtos:** Son una combinación de los dos datasets anteriores en donde se presentan datos de tipo numérico y tipo categórico. El análisis de datasets con tipos de datos mixtos ha comenzado a tomar gran interés ya que en aplicaciones de la vida real los datasets con atributos de tipo mixto son muy comunes. Al utilizar algún dataset con datos mixtos, se tenía la problemática de convertir variables categóricas a numéricas o viceversa.

(Stanford University)

V. METODOLOGÍA

SNAP (Stanford Nred Análisis Plataform):

La recolección de los datos puede tomar una cantidad considerable de tiempo; por tal motivo simplificaremos la tarea enfocándonos en un subconjunto de datos o usando un conjunto de datos ya existente.

(Stanford University)

API AWS (Amazon Web Services): Con Amazon Elastic MapReduce (Amazon EMR) una herramienta de AWS; se puede analizar y procesar grandes cantidades de datos. Lo hace mediante la distribución del trabajo de cómputo en un clúster de servidores virtuales que se ejecutan en la nube de Amazon. El grupo se gestiona mediante un marco de código abierto llamado Hadoop.

(Amazon Web Service)

El concepto MapReduce lo introdujo Google en 2004 en el paper "MapReduce: Simplified Data Processing on Large Clusters".

El objetivo principal de MapReduce era permitir la computación paralela sobre grandes colecciones

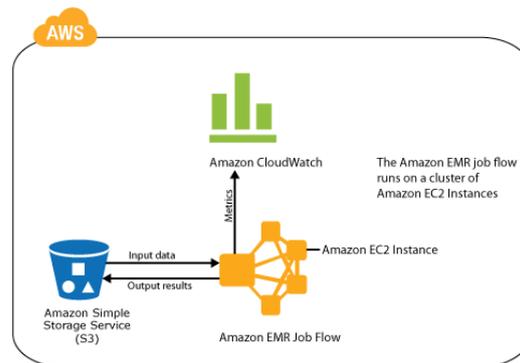


Figura 6. AEMR

de datos permitiendo abstraerse de los grandes problemas de la computación distribuida.

MapReduce consta de 2 fases: Map y Reduce. Las funciones Map y Reduce se aplican sobre pares de datos (clave, valor).

- "Map": Toma como entrada un par (clave, valor) y devuelve una lista de pares (clave2, valor2). Esta operación se realiza en paralelo para cada par de datos de entrada. Luego el framework MapReduce (como Hadoop MapReduce) agrupa todos los pares generados con la misma clave de todas las listas, creando una lista por cada una de las claves generadas.
- "Reduce": Se realiza en paralelo tomando como entrada cada lista de las obtenidas en el Map y produciendo una colección de valores. (Amazon Web Service)

MACHINE LEARNING Machine Learning se enfoca en el diseño de soluciones de problemas con aplicaciones que incluyen motores de búsqueda. En nuestro caso el proyecto que resolveremos con la ayuda de Machine Learning podemos extraer valores de la inmensa fuente de datos que poseemos y mientras más datos tenga el algoritmo va a ser más preciso el desarrollo.

El Machine Learning se divide en dos áreas principales: aprendizaje supervisado y aprendizaje no supervisado. Aunque pueda parecer que el primero se refiere a la predicción con intervención

humana y la segunda no, estos dos conceptos tienen más que ver con qué queremos hacer con los datos.

- El objetivo del **aprendizaje supervisado** es hacer predicciones a futuro basadas en comportamientos o características que se han visto en los datos ya almacenados (el histórico de datos). El aprendizaje supervisado permite buscar patrones en datos históricos relacionando los todos campos con el campo objetivo. Por ejemplo, los correos electrónicos se etiquetan como “spam” o “legítimo” por parte de los usuarios. El proceso de predicción se inicia con un análisis de qué características o patrones tienen los correos ya marcados con ambas etiquetas. Se puede determinar, por ejemplo, que un correo spam es aquel que viene de determinadas direcciones IP, y además tiene una determinada relación texto/imágenes, y además contiene ciertas palabras, y además no hay nadie en el campo “Para:”, y además... Este sería tan solo uno de los patrones. Una vez determinados todos los patrones (esta fase se llama “de aprendizaje”), los correos nuevos que nunca han sido marcados como spam o legítimos se comparan con los patrones y se clasifican (se predice) como “spam” o legítimos” en función de sus características.

- Por otro lado, el **aprendizaje no supervisado** usa datos históricos que no están etiquetados. El fin es explorarlos para encontrar alguna estructura o forma de organizarlos. Por ejemplo, es frecuente su uso para agrupar clientes con características o comportamientos similares a los que hacer campañas de marketing altamente segmentadas.

(Departamento de Ciencias de la Computacion e Inteligencia Artificial)

TIPOS DE CLUSTERING Los algoritmos de agrupación de clustering varían entre sí por las reglas heurísticas que utilizan y el tipo de aplicación para el cual fueron diseñados. La mayoría de ellos se basa en el empleo sistemático de distancias entre vectores (objetos a agrupar) así como entre clústers o grupos que se van formando a lo largo del proceso de clustering.

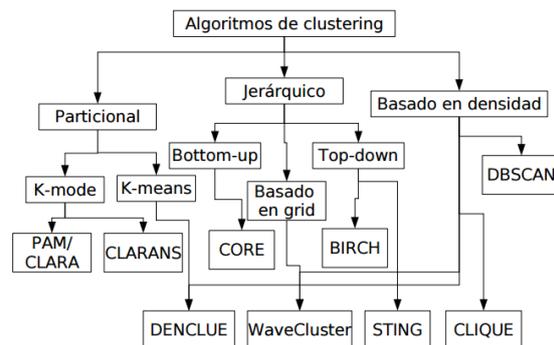


Figura 7. Técnicas de Clustering

Los **Algoritmos Jerárquicos** producen una secuencia anidada de particiones del conjunto de objetos, es decir, los grupos se organizan de forma jerárquica y cada grupo (cluster) puede verse como la unión de otros grupos (clusters), obteniendo así distintos niveles de jerarquía de grupos. Esta organización jerárquica es representada tradicionalmente por un árbol llamado dendrograma, el cual proporciona una taxonomía o índice jerárquico de la información procesada.

(tesisEdnaHernandez.pdf)

Los métodos Jerárquicos crean una descomposición jerarquía del conjunto de datos. Estos métodos pueden ser calificados como aglomerativos o divisivo, basado en cómo se comporta la descomposición, esta puede ser de abajo-arriba (bottom-up, merging) o de arriba-abajo (top-down, splitting). La calidad de los métodos jerárquicos sufre puesto que una vez que se ha dado una división o una unión, no se puede reajustar, ósea hay que reiniciar el proceso.

(tesisEdnaHernandez.pdf)

VI. RESULTADOS

Nosotros usamos el tipo de clustering jerárquico divisivo porque tenemos un solo dataset de productos de amazon y lo que queremos determinar son las categorías de los productos. Con este estudio que hemos realizado esperamos que las herramientas utilizadas sean de ayuda a nuevos

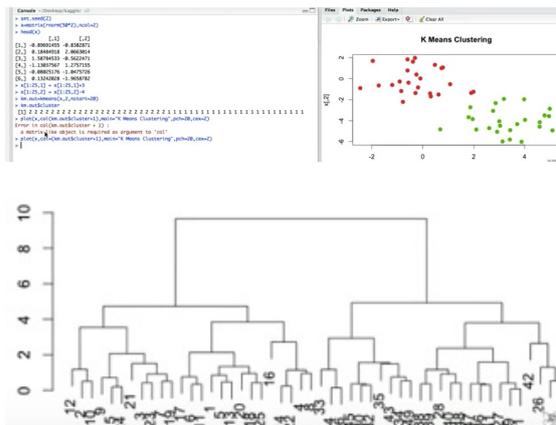


Figura 8. Dendrograma del Cluster Jerarquico

estudios o resoluciones de problemáticas similares.

Con el uso de las herramientas expuestas en este artículo hemos obtenido los diferentes clusters generados por el algoritmo jerárquico utilizado mostrando los grupos categóricos de cada producto, dando la mejor opción al cliente al momento de elegir un artículo con la información y características necesarias desechando aquellos datos irrelevantes. Con la precisión que maneja el algoritmo jerárquico, básicamente lo que obtenemos son varios grupos de productos unidos por sus características semejantes formando el cluster de forma jerárquica.

El cluster jerárquico utilizado nos agrupa varios artículos y según la cantidad de datos y las características similares entre ellos va lanzando los diferentes cluster hasta llegar al agrupamiento total. Como observamos en la imagen tenemos un cluster que a su vez contiene varios clusters que se van agrupando de forma jerárquica.

VII. CONCLUSIONES

Todas las herramientas utilizadas en este proyecto nos han ayudado y nos han enseñado la manipulación de grandes cantidades de datos y a su vez hemos resuelto un problema que es de mucha ayuda para todos aquellos que compran en la página de Amazon ya que serían más los beneficiados al ver que cuando busquen un artículo

se mostrará el de mejor calidad. Nosotros pudimos notar que la minería de datos hoy en día es muy importante y que haber trabajado en este proyecto nos dejó un aprendizaje importante con respecto a las herramientas utilizadas.

REFERENCIAS

- [1]@bookpena2002analysis, title=Análisis de datos multivariantes, author=Peña, Daniel, volume=24, year=2002, publisher=McGraw-Hill Madrid
- [2]@articleguide2010amazon, title=Amazon Elastic MapReduce, author=Guide, Developer, year=2010
- [3]@articledean2008mapreduce, title=MapReduce: simplified data processing on large clusters, author=Dean, Jeffrey and Ghemawat, Sanjay, journal=Communications of the ACM, volume=51, number=1, pages=107–113, year=2008, publisher=ACM
- [4]@booklopez2007minería, title=Minería de datos: técnicas y herramientas, author=López, César Pérez, year=2007, publisher=Editorial Paraninfo
- [5]@articlefernandez1991analysis, title=El análisis de cluster: aplicación, interpretación y validación, author=Fernández Santana, Óscar, journal=Papers: revista de sociología, number=37, pages=065–76, year=1991
- [6]@articlegarciaibig, title=Big Data: Diseño de algoritmos para clasificación extremadamente no balanceada, author=García, Sara Del Río and Sánchez, José Manuel Benítez
- [7]@articlekanungo2002efficient, title=An efficient k-means clustering algorithm: Analysis and implementation, author=Kanungo, Tapas and Mount, David M and Netanyahu, Nathan S and Piatko, Christine D and Silverman, Ruth and Wu, Angela Y, journal=Pattern Analysis and Machine Intelligence, IEEE Transactions on, volume=24, number=7, pages=881–892, year=2002, publisher=IEEE
- [8]@inproceedingswagstaff2001constrained, title=Constrained k-means clustering with background knowledge, author=Wagstaff, Kiri and Cardie, Claire and Rogers, Seth and Schrödl, Stefan and others, booktitle=ICML, volume=1, pages=577–584, year=2001