ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL.

Facultad de Ingeniería en Estadística Informática

"AGRUPACION AUTOMATICA DE DESCRIPCIONES DE PRODUCTOS"

TESIS DE GRADO

Previa a la obtención del Título de:

INGENIERO EN ESTADISTICA INFORMATICA

Presentado por

Luis Enrique Loor Díaz

Guayaquil - Ecuador

2006

AGRADECIMIENTO

A todas las personas que de una u otra forma colaboraron en la realización de este trabajo especialmente a mis padres y a Dios que siempre me brindaron ese apoyo incondicional.

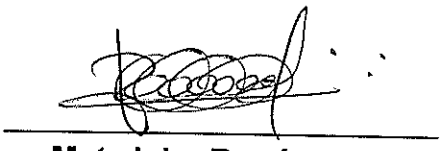
DEDICATORIA

A MIS PADRES, A MIS HERMANOS

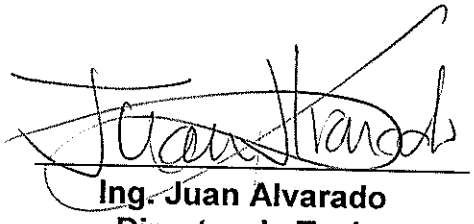
Y A ESA PERSONA INCONDICIONAL

QUE ESTUVO AHÍ SIEMPRE….

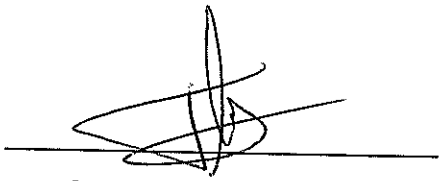
TRIBUNAL DE GRADUACIÓN



Mat. John Ramírez
Presidente



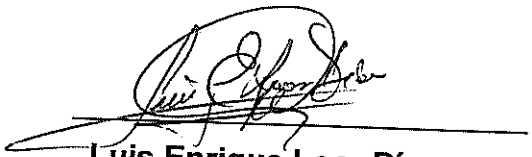
Ing. Juan Alvarado
Director de Tesis



Ing. Soraya Solís
Vocal

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de grado, me corresponden exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITECNICA DEL LITORAL"



Luis Enrique Loor Díaz

INDICE

RESUMEN

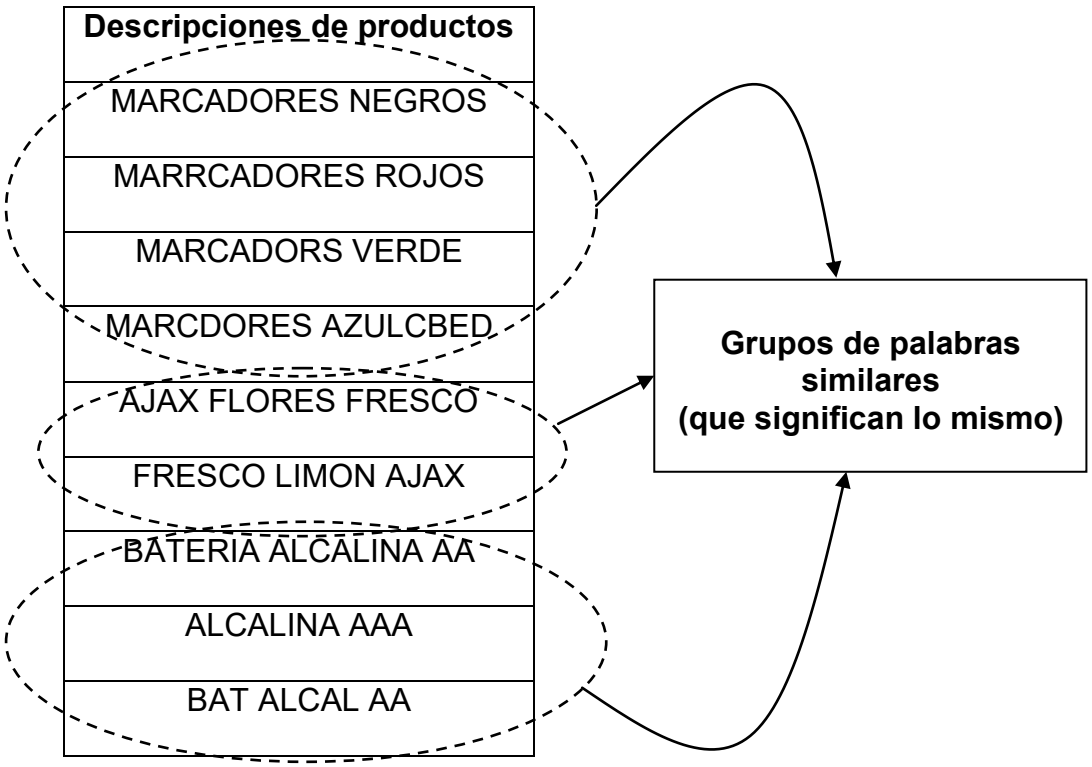
En este trabajo se realiza una aplicación acerca de la utilización del Data Mining. El propósito de esta tesis es automatizar el proceso de agrupamiento de las descripciones de productos de una gran cantidad de registros, para así tener grupos que puedan ser codificados posteriormente y así realizar inferencias sobre estos grupos que son una cantidad menor de datos, y que a la vez sea representativa.

En este trabajo se considera que, se toma una muestra de la población total (combinaciones de palabras), para poder realizar el análisis, ya que si tomamos la población total aunque serían más confiables los resultados no sería óptimo debido a que realizamos una comparación secuencial de cadenas de caracteres.

Cabe recalcar que esta tesis es una parte importante del proceso de extracción del conocimiento, porque permiten la agrupación de

registros que antes se presentaban como separados (descripciones de productos) en cadenas de caracteres y esta agrupación nos permite un procesamiento de los datos de estos productos independientemente en que forma son presentados en las descripciones de estos productos.

Este problema se lo conoce como Data Matching o De Duplex.



INTRODUCCIÓN

En el campo del Reconocimiento de formas o patrones, las técnicas de clasificación basadas en distancia, necesitan de la obtención de prototipos adecuados para cada clase. Una de las posibilidades es usar la media de la clase (o el conjunto formado por la media de las diversas subclases que componen la clase) como prototipo de la misma.

Cuando se habla de espacios euclídeos (representación vectorial), hallar la media es un problema sencillo, pero no así si usamos la representación por cadenas de caracteres. En dicho caso, el problema de hallar la cadena media es duro ya que no existe el concepto de promedio de cadenas.
Así, se pasan a definir aproximaciones sobre la cadena media para dichos usos.

En este trabajo se propone usar la distancia de Levenshtein para obtener un valor por medio del cual mediremos la similaridad entre cadenas y así agrupar las descripciones, lo cual definiremos más adelante.

Objetivo de la Tesis

El objetivo de esta Tesis de grado es:

Crear una aplicación para la agrupación automática de las descripciones de productos.

Facilitar a las Organizaciones que manejan grandes bases de datos (millones de registros) la codificación de los registros de manera categórica.

Definición del Problema

Las organizaciones y empresas en general que manejan bases de datos extensas, tienen inconvenientes el momento de tener sus registros codificados de una manera categórica cuando de descripciones o detalles se trata. De aquí nace la necesidad, como para el Servicio de Aduanas, de crear una aplicación que agrupe las descripciones que están expresadas en cadenas de caracteres. ¿Por qué?, porque se desea unificar todas las descripciones que representen lo mismo para el correcto procesamiento de los grupos de descripciones de productos para ser aplicados en problemas como el procesamiento de los precios de las mercancías importadas y generar estimados del precio de dichas mercaderías importadas, problema que es resuelto finalmente con técnicas de minerías de datos.

CAPÍTULO 1

1. DATA MINING

1.1 Generalidades

Las técnicas del Data Mining son el resultado de un largo proceso de investigación. El Data Mining, es la exploración y análisis de grandes cantidades de datos para descubrir modelos significantes y reglas. También está dirigido a explicar o categorizar algún campo designado particular como detalles, descripciones, etc.

El Data Mining está principalmente dirigido a construir modelos entre los grandes grupos de datos o archivos. Un modelo Data Mining simplemente es un algoritmo o juego de reglas que conectan datos a un blanco particular o resultado.

Bajo las circunstancias correctas, un modelo puede producir una visión proporcionando una explicación de los resultados de un interés

particular, como hacer un pedido o una orden de compra con un código que represente un determinado grupo de productos.

Data Mining está listo para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

- Recolección masiva de datos
- Potentes computadoras con multiprocesadores
- Algoritmos de Data Mining

Los registros de las bases de datos comerciales están creciendo a un ritmo sin precedentes. Un reciente estudio del META GROUP sobre los proyectos de Data Warehouse encontró que el 19% de los que contestaron están por encima del nivel de los 50 Gigabytes, mientras que el 59% espera alcanzarlo en el segundo trimestre de 2006. En algunas industrias, tales como ventas al por menor (retail), estos números pueden ser aún mayores. MCI Telecommunications Corp. cuenta con una base de datos de 3 terabytes + 1 terabyte de índices y overhead corriendo en MVS sobre IBM SP2. La necesidad paralela de motores computacionales mejorados puede ahora alcanzarse de forma más costo - efectiva con tecnología de computadoras con multiprocesamiento paralelo. Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace 10 años, pero que

sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que consistentemente son más performantes que métodos estadísticos clásicos.

En la evolución desde los datos de negocios a información de negocios, cada nuevo paso se basa en el previo. Por ejemplo, el acceso a datos dinámicos es crítico para las aplicaciones de navegación de datos (drill through applications), y la habilidad para almacenar grandes bases de datos es crítica para Data Mining.

Los componentes esenciales de la tecnología de Data Mining han estado bajo desarrollo por décadas, en áreas de investigación como estadísticas, inteligencia artificial y aprendizaje de máquinas. Hoy, la madurez de estas técnicas, junto con los motores de bases de datos relacionales de alta performance, hicieron que estas tecnologías fueran prácticas para los entornos de data warehouse actuales.

1.2 ¿Qué es Data Mining?

Data Mining, la extracción de información oculta y predecible de grandes bases de datos, es una poderosa tecnología nueva con gran potencial que ayuda a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse).

Un Sistema Datamining es una tecnología de soporte para usuario final cuyo objetivo es extraer conocimiento útil y utilizable a partir de la información contenida en las bases de datos de las empresas.

1.2.1 Desarrollo de los sistemas Data Mining

Los sistemas Datamining se desarrollan bajo lenguajes de última generación basados en la inteligencia artificial y utilizan modelos matemáticos tales como:

- **Redes neuronales artificiales:** modelos predecibles no-lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.
- **Arboles de decisión:** estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Métodos específicos de árboles de decisión incluyen Arboles de Clasificación y Regresión (CART: Classification And Regression Tree) y Detección de Interacción Automática de Chi Cuadrado (CHAI: Chi Square Automatic Interaction Detection)
- **Algoritmos genéticos:** técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y

selección natural en un diseño basado en los conceptos de evolución.

- **Método del vecino más cercano:** una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases del/de los k registro (s) más similar/es a él en un conjunto de datos históricos (donde k ? 1). Algunas veces se llama la técnica del vecino k-más cercano.
- **Regla de inducción:** la extracción de reglas if-then de datos basados en significado estadístico.

Muchas de estas tecnologías han estado en uso por más de una década en herramientas de análisis especializadas que trabajan con volúmenes de datos relativamente pequeños. Estas capacidades están ahora evolucionando para integrarse directamente con herramientas OLAP y de Data Warehousing. [ref.1]

1.2.2 ¿Que hacen las herramientas DataMining?

Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información. Los análisis automatizados ofrecidos por el Data Mining

van más allá de los eventos pasados provistos por herramientas típicas de sistemas de soporte de decisión.

Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas y a los cuales los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

Una vez que las herramientas de Data Mining fueron implementadas en computadoras cliente servidor de alto performance o de procesamiento paralelo, pueden analizar bases de datos masivas para brindar respuesta a preguntas tales como, "¿Cuáles clientes tienen más probabilidad de responder al próximo mailing promocional, y por qué? y presentar los resultados en formas de tablas, con gráficos, reportes, texto, hipertexto, etc.

Las técnicas de Data Mining pueden ser implementadas rápidamente en plataformas ya existentes de software y hardware para acrecentar el valor de las fuentes de información existentes y pueden ser integradas con nuevos productos y sistemas pues son traídas en línea (on-line).

1.2.3 El Alcance del Data Mining

Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer estas capacidades:

- ✓ Predicción automatizada de tendencias y comportamientos.
- ✓ Data Mining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos. Un típico ejemplo de problema predecible es el marketing apuntado a objetivos (targeted marketing). Data Mining usa datos en mailing promocionales anteriores para identificar posibles objetivos para maximizar los resultados de la inversión en futuros mailing. Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.
- ✓ Descubrimiento automatizado de modelos previamente desconocidos.
- ✓ Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso.

Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar datos anormales que pueden representar errores de tipeado en la carga de datos.

- ✓ Las técnicas de Data Mining pueden redituar los beneficios de automatización en las plataformas de hardware y software existentes y puede ser implementadas en sistemas nuevos a medida que las plataformas existentes se actualicen y nuevos productos sean desarrollados. Cuando las herramientas de Data Mining son implementadas en sistemas de procesamiento paralelo de alto performance, pueden analizar bases de datos masivas en minutos. Procesamiento más rápido significa que los usuarios pueden automáticamente experimentar con más modelos para entender datos complejos. Alta velocidad hace que sea práctico para los usuarios analizar inmensas cantidades de datos. Grandes bases de datos, a su vez, producen mejores predicciones.

Las bases de datos pueden ser grandes tanto en profundidad como en ancho:

- ✓ Más columnas. Los analistas muchas veces deben limitar el número de variables a examinar cuando realizan análisis

manuales debido a limitaciones de tiempo. Sin embargo, variables que son descartadas porque parecen sin importancia pueden proveer información acerca de modelos desconocidos. Un Data Mining de alto rendimiento permite a los usuarios explorar toda la base de datos, sin preseleccionar un subconjunto de variables.

- ✓ Más filas. Muestras mayores producen menos errores de estimación y desvíos, y permite a los usuarios hacer inferencias acerca de pequeños pero importantes segmentos de población.

1.2.4 ¿Cómo Trabaja el Data Mining?

¿Cuán exactamente es capaz Data Mining de decirle cosas importantes que usted desconoce o que van a pasar? La técnica usada para realizar estas hazañas en Data Mining se llama Modelado. Modelado es simplemente el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta. Por ejemplo, si busca un galeón español hundido en los mares lo primero que podría hacer es investigar otros tesoros españoles que ya fueron encontrados en el pasado. Notaría que esos barcos frecuentemente fueron encontrados fuera de las costas de Bermuda y que hay ciertas

características respecto de las corrientes oceánicas y ciertas rutas que probablemente tomara el capitán del barco en esa época. Usted nota esas similitudes y arma un modelo que incluye las características comunes a todos los sitios de estos tesoros hundidos. Con estos modelos en mano sale a buscar el tesoro donde el modelo indica que en el pasado hubo más probabilidad de darse una situación similar. Con un poco de esperanza, si tiene un buen modelo, probablemente encontrará el tesoro.

Este acto de construcción de un modelo es algo que la gente ha estado haciendo desde hace mucho tiempo, seguramente desde antes del auge de las computadoras y de la tecnología de Data Mining. Lo que ocurre en las computadoras, no es muy diferente de la manera en que la gente construye modelos. Las computadoras son cargadas con mucha información acerca de una variedad de situaciones donde una respuesta es conocida y luego el software de Data Mining en la computadora debe correr a través de los datos y distinguir las características de los datos que llevarán al modelo. Una vez que el modelo se construyó, puede ser usado en situaciones similares donde usted no conoce la respuesta.

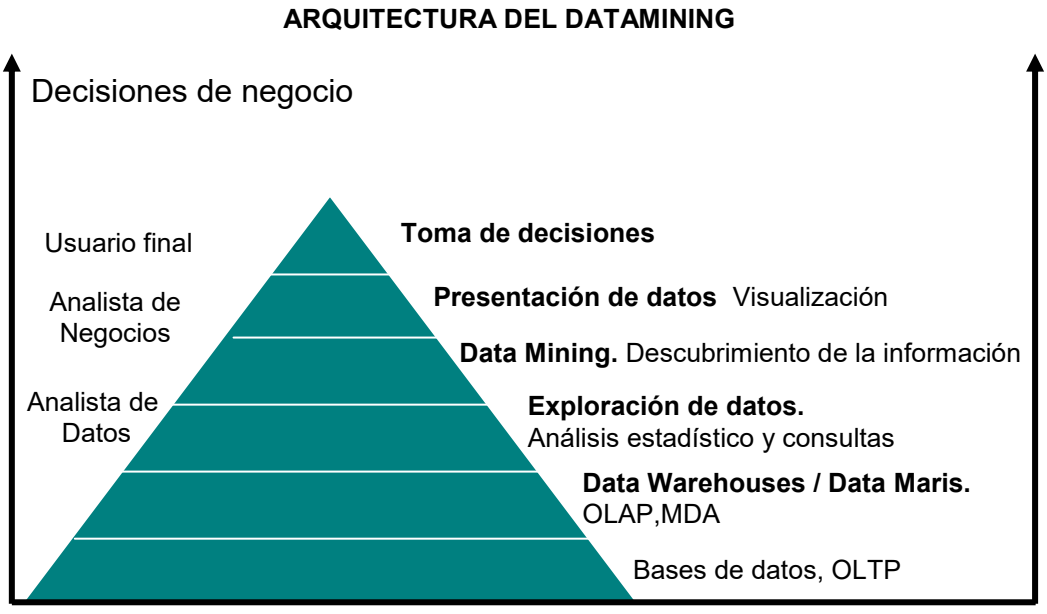
Si alguien le dice que tiene un modelo que puede predecir el uso de los clientes, ¿Cómo puede saber si es realmente un buen modelo? La

primera cosa que puede probar es pedirle que aplique el modelo a su base de clientes - donde usted ya conoce la respuesta. Con Data Mining, la mejor manera para realizar esto es dejando de lado ciertos datos para aislarlos del proceso de Data Mining. Una vez que el proceso está completo, los resultados pueden ser testeados contra los datos excluidos para confirmar la validez del modelo. Si el modelo funciona, las observaciones deben mantenerse para los datos excluidos.

1.3 Arquitectura para Data Mining

[ref. 1] Para aplicar mejor estas técnicas avanzadas, éstas deben estar totalmente integradas con el data warehouse así como con herramientas flexibles e interactivas para el análisis de negocios.

Figura 1.1



Varias herramientas de Data Mining actualmente operan fuera del warehouse, requiriendo pasos extra para extraer, importar y analizar los datos. Además, cuando nuevos conceptos requieren implementación operacional, la integración con el warehouse simplifica la aplicación de los resultados desde Data Mining. El Data warehouse analítico resultante puede ser aplicado para mejorar procesos de negocios en toda la organización, en áreas tales como manejo de campañas promocionales, detección de fraudes, lanzamiento de nuevos productos, etc.

El punto de inicio ideal es un data warehouse que contenga una combinación de datos de seguimiento interno de todos los clientes junto con datos externos de mercado acerca de la actividad de los competidores. Información histórica sobre potenciales clientes también provee una excelente base para prospecting. Este warehouse puede ser implementado en una variedad de sistemas de bases relacionales y debe ser optimizado para un acceso a los datos flexible y rápido.

Un server multidimensional OLAP permite que un modelo de negocios más sofisticado pueda ser aplicado cuando se navega por el data warehouse. Las estructuras multidimensionales permiten que el usuario analice los datos de acuerdo a como quiera mirar el negocio -

resumido por línea de producto, u otras perspectivas claves para su negocio. El server de Data Mining debe estar integrado con el data warehouse y el server OLAP para insertar el análisis de negocios directamente en esta infraestructura. Un avanzado, metadata centrado en procesos define los objetivos del Data Mining para resultados específicos tales como manejos de campaña, prospecting, y optimización de promociones. La integración con el data warehouse permite que decisiones operacionales sean implementadas directamente y monitoreadas. A medida que el data warehouse crece con nuevas decisiones y resultados, la organización puede "minar" las mejores prácticas y aplicarlas en futuras decisiones.

Este diseño representa una transferencia fundamental desde los sistemas de soporte de decisión convencionales. Más que simplemente proveer datos a los usuarios finales a través de software de consultas y reportes, el server de Análisis Avanzado aplica los modelos de negocios del usuario directamente al warehouse y devuelve un análisis proactivo de la información más relevante. Estos resultados mejoran los metadatos en el server OLAP proveyendo una estrato de metadatos que representa una vista fraccionada de los datos. Generadores de reportes, visualizadores y otras herramientas de análisis pueden ser aplicadas para planificar futuras acciones y confirmar el impacto de esos planes.

1.4 ¿Por que usar Data Mining?

Si bien es cierto data mining se presenta como una tecnología emergente, posee ciertas ventajas, como ser:

- Resulta un buen punto de encuentro entre los investigadores y las personas de negocios.
- Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios.
- Trabajar con esta tecnología implica cuidar un sin número de detalles debido a que el producto final involucra "toma de decisiones".
- Contribuye a la toma de decisiones tácticas y estratégicas proporcionando un sentido automatizado para identificar información clave desde volúmenes de datos generados por procesos tradicionales y de e-Business.
- Permite a los usuarios dar prioridad a decisiones y acciones mostrando factores que tienen un mayor en un objetivo, qué segmentos de clientes son desechables y qué unidades de negocio son sobrepasados y por qué.

- Proporciona poderes de decisión a los usuarios del negocio que mejor entienden el problema y el entorno y es capaz de medir la acciones y los resultados de la mejor forma.
- Genera Modelos descriptivos: en un contexto de objetivos definidos en los negocios permite a empresas, sin tener en cuenta la industria o el tamaño, explorar automáticamente, visualizar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales de la cuenta de resultados (tales como el aumento de los ingresos, incremento de los beneficios, contención de costes y gestión de riesgos).
- Genera Modelos predictivos: permite que relaciones no descubiertas e identificadas a través del proceso del Data Mining sean expresadas como reglas de negocio o modelos predictivos. Estos outputs pueden comunicarse en formatos tradicionales (presentaciones, informes, información electrónica compartida, embebidos en aplicaciones, etc.) para guiar la estrategia y planificación de la empresa.

1.5 Data Mining versus Estadística

- La diferencia decisiva entre Data Mining y Estadística es la dirección de la búsqueda (query).
- En el Data Mining, la interrogación de los datos se hace mediante algoritmos de Inteligencia Artificial (desde ahora, IA) o Redes Neuronales, en vez de a partir de la contribución del estadístico o del análisis de negocios.
- En otras palabras, el Data Mining está dirigido por la naturaleza de los datos, en vez de estar dirigido por el usuario o por la verificación, como ocurre en la mayoría de los análisis estadísticos.
- **Data Mining** tiene también grandes ventajas sobre la Estadística cuando la escala de las bases de datos aumenta de tamaño, simplemente porque los enfoques manuales del análisis de datos se están haciendo impracticables.

1.6 Metodología del Data Mining

La metodología del Data Mining tiene 11 pasos:

1.- Traducir el problema comercial o de negocios en un problema de minería de datos.

2.- Seleccionar los datos apropiados.

3.- Saber conseguir los datos.

4.- Crear un modelo ejemplar.

5.- Arreglar problemas con los datos.

6.- Transformar los datos para traer información.

7.- Modelos Robustos.

8.- Modelos de Evaluación.

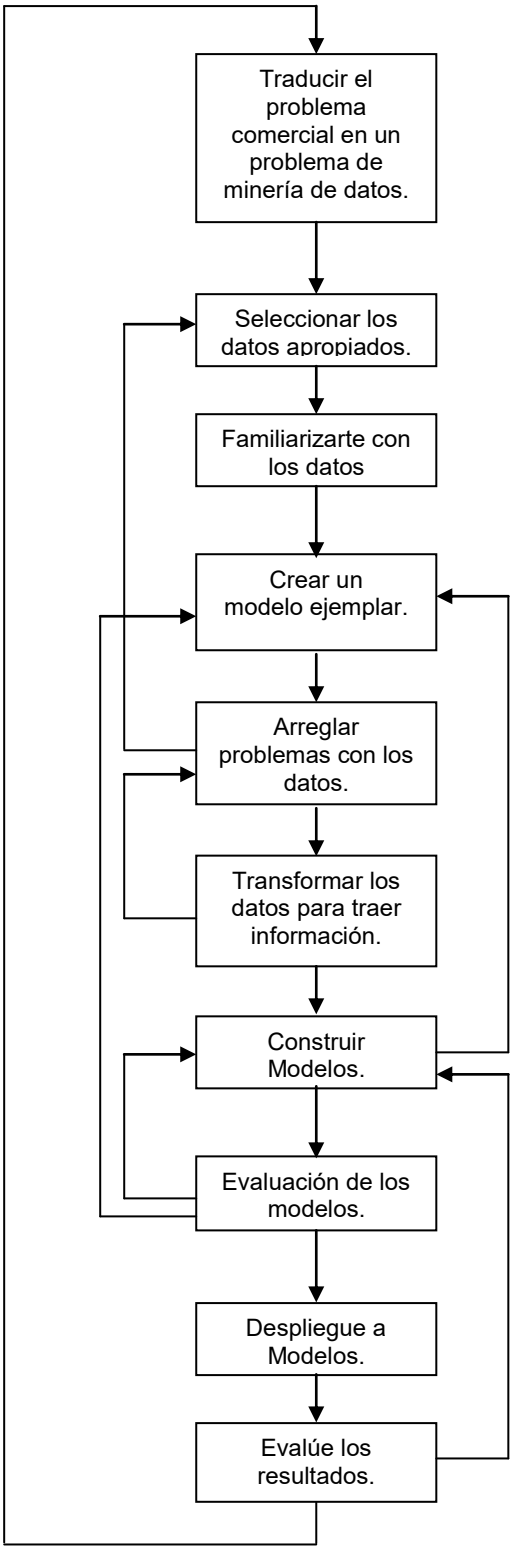
9.- Despliegue a Modelos.

10.- Evalúe los resultados.

11.- Empiece de nuevo.

Como se muestra en la siguiente figura, el proceso del Data Mining es un conjunto de vueltas anidadas interactivas en lugar de una línea recta. Los pasos tienen un orden natural pero no es necesario seguirlo para terminar con un paso antes del siguiente.

PROCESO DEL DATA MINING



Paso 1: Traducir el problema comercial en un problema de minería de datos.

Para transformar un problema de negocios en un problema de minería de datos, este debe ser reformulado como una de las seis tareas del data mining, tales como: clasificación, estimación, predicción, afinidad de agrupamiento, clustering, descripción y perfilamiento.

La tarea del modelamiento es encontrar reglas que expliquen los valores conocidos de las variables en blanco.

Paso 2: Seleccionar los datos apropiados

La extracción del conocimiento requiere datos. Los datos requeridos estarían en un almacén corporativo de los datos, limpiados, disponible, históricamente exacto, y puestos al día con frecuencia. De hecho, se dispersa más a menudo en una variedad de sistemas operacionales en formatos incompatibles en la computadora, que funcionan diversos sistemas operativos, alcanzada a través de las herramientas de escritorio incompatibles.

Las fuentes de datos que son útiles y disponibles varían, por supuesto, de problema en problema y de industria en industria. Algunos ejemplos de datos útiles son:

- ✓ La garantía demanda datos (los campos incluyendo de fijo-formato y del texto libre).
- ✓ Expedientes de las cargas de tarjetas de crédito.
- ✓ Datos de puntos de venta (incluyendo códigos de anillo, cupones, descuentos aplicados)
- ✓ El seguro médico demanda de datos.
- ✓ Datos de registro del Web.
- ✓ Registros del uso del servidor del comercio electrónico.
- ✓ Expedientes de la respuesta del correo directo.
- ✓ Expedientes de centros de llamadas, incluyendo las notas escritas por los responsables del centro de llamadas.
- ✓ Expedientes del funcionamiento de la prensa.
- ✓ Expedientes de registro de motores de vehículos.
- ✓ Nivel de ruido en decibeles de micrófonos ubicados en comunidades cerca del aeropuerto.
- ✓ Expedientes de detalles de llamadas telefónicas.
- ✓ Datos de la respuesta del examen.
- ✓ Datos demográficos y de estilo de vida.
- ✓ Datos económicos.

Una vez que el problema de negocios ha sido formulado, es posible formar una lista deseada de datos que serían útiles tener. El primer paso es

buscar los datos disponibles y crear una lista de candidatos a problemas de negocios.

Paso 3: Familiarizarse con los datos

Los buenos mineros de datos se parecen confiar mucho en la intuición de alguna manera que puede conjeturar lo que pudo ser una buena variable derivada a intentar, por ejemplo. La única manera de desarrollar la intuición para conocer lo que está pasando en un grupo de datos desconocido, es sumergirse en él. De manera que a la larga, se haga probable descubrir muchos problemas de la calidad de los datos, para responder cualquier pregunta que surja en determinado momento.

Paso 4: Crear un grupo de datos

El sistema del modelo contiene todos los datos que son usados en el proceso del modelo. Algunos de los datos son usados para encontrar patrones y para verificar si el sistema del modelo de datos creado, es estable. Crear un sistema del modelo requiere datos que provienen de fuentes múltiples para el formulario de firmas del cliente y después preparar los datos para el análisis.

Paso 5: Reparar problemas con los datos

Todos los datos son sucios. Todos los datos tienen problemas. Cuál es o no es un problema varía con la técnica de extracción del conocimiento. Para algunos, tal como árboles de decisión, los valores que faltan, y los afloramientos no causan demasiado apuro. Para otras, tales como redes neuronales, causan toda clase de apuro. Por esa razón, algunos de lo que tenemos que decir sobre problemas que fijan con datos se pueden encontrar en los capítulos en las técnicas donde causan la mayoría de la dificultad.

A continuación se nombran los problemas más comunes que necesitan ser reparados:

- ✓ Variables categóricas con demasiados valores.
- ✓ Variables numéricas con distribuciones sesgadas y outliers.
- ✓ Valores faltantes.
- ✓ Valores con los significados que cambian en un cierto plazo.
- ✓ Codificación contraria de los datos.

Paso 6: Transformar los datos para traer información.

Una vez que los datos hayan sido revisados y los problemas importantes de los datos han sido reparados, los datos se deben preparar aún para el

análisis. Esto implica el agregar campos derivados para traer la información a la superficie. Puede también implicar el quitar los outliers (puntuaciones extremas dentro una variable), variables numéricas, el agrupar clases para las variables categóricas, aplicando transformaciones tales como logaritmos, el dar vuelta de cuenta a proporciones, y los similares.

He aquí algunos ejemplos de transformaciones:

- ✓ Las tendencias de la captura.
- ✓ Crear cocientes y otras combinaciones de las variables.
- ✓ Conversión de cuentas a proporciones.

Paso 7: Construir modelos

Los detalles de este paso varían de técnica en técnica y se describen en los capítulos dedicados a cada método de minería de datos. De modo general, éste es el paso donde la mayor parte del trabajo ocurre en crear un modelo. En la explotación minera dirigida a datos, el sistema de entrenamiento se utiliza para generar una explicación de la variable

independiente o variable en blanco en términos de independiente o variables de entrada.

Esta explicación puede tomar la forma de una red neuronal, de un árbol de la decisión, de un gráfico del acoplamiento, o de una cierta otra representación de la relación entre la variable blanco y los otros campos en la base de datos. En la minería de datos indirecta, no hay variable blanco. El modelo encuentra relaciones entre los registros y los expresa como reglas de asociación o asignándolas a los clusters comunes.

La construcción de modelos es el único paso del proceso minería de datos que ha sido automatizado verdaderamente un moderno software de minería de datos. Por esa razón, relativamente toma poco tiempo un proyecto de minería de datos.

Paso 8: Evaluación de los modelos

En este paso se determina si los modelos están trabajando o no. Un modelo determinado debe responder preguntas tales como:

¿Qué tan exacto es el modelo?

¿Cuan bien el modelo describe los datos observados?

¿Cuánta confianza se puede poner en las predicciones del modelo?

¿Cuán comprensible es el modelo?

Por supuesto, las respuestas a este tipo de preguntas corresponden al tipo del modelo que fue construido. Lo que refiere básicamente es a los méritos técnicos del modelo.

Paso 9: Despliegue de los modelos

Desplegar un modelo significa el movimiento que hay de él, desde el ambiente de minería de datos al ambiente que llega. Este proceso puede ser fácil o difícil. En el peor de los casos, el modelo se desarrolla en un ambiente especial usando un software que no funciona en ninguna parte.

Para desplegar el modelo, un programador toma una descripción impresa del modelo y la recodifica en otro lenguaje de programación de tal forma que pueda funcionar en la plataforma a la que llega.

Un problema más común es que el modelo utiliza las variables de entrada que no están en los datos originales. Esto no debe ser un problema puesto que las entradas se derivan por lo menos de campos que fueron extraídos originalmente del sistema del modelo.

Desafortunadamente, los mineros de datos no son siempre buenos al guardar un registro limpio, reutilizable de las transformaciones que se aplicaron a los datos.

Las cuentas representan a menudo una probabilidad que son valores típicamente numéricos entre 0 y 1, pero de ninguna manera tan necesariamente. Una cuenta pudo también ser una etiqueta de la clase proporcionada por un modelo de clustering, por ejemplo, o una etiqueta de la clase de una probabilidad.

Paso 10: Determinar los resultados

Este paso nos ayuda a la toma de decisiones en la parte final del proceso del Data Mining. Una gráfica útil demostraría cuántos dólares se traen adentro para un gasto dado en la campaña de la comercialización. Después de todo, si desarrollar el modelo es muy costoso, un correo de la masa puede ser más rentable que el anterior nombrado.

¿Cuál es el coste fijo de creación de la campaña y el modelo que la ayuda?

¿Cuál es el coste por el recipiente de hacer la oferta?

¿Cuál es el coste por la respuesta de satisfacer la oferta?

¿Cuál es el valor de una respuesta positiva?

En fin, la medida que cuenta es la vuelta en la inversión. La elevación que mide un sistema ayuda en la prueba de elegir el modelo correcto. Pero, es muy importante medir estas cosas en el campo también. En una comercialización de la base de datos el uso, éste requiere siempre poner a grupos de control a un lado y cuidadosamente seguir la respuesta del cliente según varias cuentas del modelo.

CAPITULO 2

2. Técnicas de agrupamiento

2.1 Introducción

Las técnicas de agrupamiento buscan la obtención de agrupamientos naturales dentro de un conjunto de muestras; es decir, vienen a hacer una división más o menos natural dentro del conjunto de datos disponible. Dichas técnicas son utilizadas para poder distinguir subclases dentro de una clase determinada. Un ejemplo de la necesidad que se podría presentar en un negocio es la segmentación de sus clientes.

La obtención de agrupamientos naturales responde a muy diversos criterios. En general, el factor fundamental es el uso de una medida de disimilitud entre agrupamientos. Siguiendo esta medida, lo que se

consigue es distinguir si dos muestras pertenecen al mismo agrupamiento (disimilitud baja) o no (disimilitud alta). Esta medida también se aplica agrupamiento a agrupamiento, a fin de propiciar fusiones y separaciones entre los diversos agrupamientos obtenidos. El objetivo de esto es conseguir agrupamientos lo más adecuados posibles.

Formalmente, un proceso de agrupamiento sobre un conjunto de datos X ={x1, x2,…..,xn} va a obtener c subconjuntos X1,X2,……..,Xc, Xi ≠ 0 ; para i = 1,……..,c, tales que $X = X_1 \cup X_2 \cup \ldots..\cup X_c$ y disjuntos entre sí ($Xi \cap Xj = 0$; i=j), de manera que los datos de cada Xi tengan entre sí una alta similitud.

2.2 Clustering

El término clustering básicamente significa particionar un conjunto de registros o datos en grupos de tal forma, que los datos dentro de cada grupo tengan una alta similaridad entre si, y almacenar sólo la caracterización del grupo. Podemos decir que el clustering es muy útil si el conjunto de datos está efectivamente agrupado, pero no lo es si los datos están dispersos.

Clustering que traduce también clasificación es el término utilizado para identificar las técnicas de clasificación en las cuales las clases o clusters se desconocen de antemano.

Entonces los registros son clasificados de acuerdo con la similitud entre ellos en clases no predefinidas de antemano. El problema de la minería en este caso es establecer lo que significan los grupos construidos de esta forma. Un problema de segmentación de clientes puede empezar con una tarea de clustering. La clasificación automática de documentos es otro ejemplo de clustering.

El análisis de clusters es utilizado en numerosas aplicaciones tales como reconocimiento de patrones, análisis de datos, procesamiento de imágenes e investigación de mercado. A través de la técnica de clustering pueden identificarse distintas regiones y por lo tanto descubrir patrones y relaciones interesantes entre los atributos. Como función del Data Mining (DM) , el análisis de clusters puede ser utilizado como una herramienta independiente para obtener una visión de la distribución de los datos , para observar las características de cada cluster y enfocar un análisis más exhaustivo hacia un grupo o cluster determinado.

Alternativamente, puede servir como un paso del preprocesamiento para otros algoritmos como por ejemplo para el de clasificación en el cual se trabajaría luego sobre los clusters originados. El clustering de datos está en continuo desarrollo, y debido a los grandes volúmenes de datos almacenados en las BD, el análisis de clusters se ha vuelto un tema altamente importante en los estudios de DM.

Entonces el clustering es una técnica para agrupar a los elementos de una muestra en grupos, denominados conglomerados, de tal forma que, con respecto a la distribución de los valores de las variables, por un lado, sea lo mas homogéneo posible y, por otro, sean muy distintos entre sí.

2.2.1 Requerimientos de un algoritmo de Clustering en Data Mining

Los requerimientos básicos de un algoritmo de clustering en DM son los siguientes:

- **Escalabilidad.-** La mayoría de los algoritmos de clustering trabajan de manera apropiada con un número pequeño de observaciones (hasta 200 aproximadamente), mientras que se necesita una gran escalabilidad para realizar agrupamiento de datos en bases con millones de observaciones.

- **Habilidad para trabajar con distintos tipos de atributos.-** Muchos algoritmos se han diseñado para trabajar sólo con datos numéricos, mientras que en una gran cantidad de ocasiones, es necesario trabajar con atributos asociados a tipos numéricos, binarios, discretos alfanuméricos.
- **Descubrimiento de clusters con formas arbitrarias.-** La mayoría de los algoritmos de clustering se basan en la distancia Euclídea, lo que tiende a encontrar clusters todos con forma (circular) y densidad similares. Es importante diseñar algoritmos que puedan establecer clusters de formas arbitrarias.

- **Requerimientos mínimos en el conocimiento del dominio para determinar los parámetros de entrada.-** La herramienta no debería solicitarle al usuario que introduzca la cantidad de clases que quiere considerar, ya que dichos parámetros en muchas ocasiones no son fáciles de determinar, y esto haría que sea difícil controlar la calidad del algoritmo.

- **Habilidad para tratar con datos "ruidosos".-** La mayoría de las BD contienen datos con comportamiento extraño, datos faltantes, desconocidos o erróneos. Algunos algoritmos de

clustering son sensibles a tales datos y pueden derivarlos a clusters de baja calidad.

- **Insensibilidad al orden de las observaciones de entrada.-** Algunos algoritmos son sensibles al orden en que se consideran las observaciones. Por ejemplo, para un mismo conjunto de datos, dependiendo del orden en que se analicen, los clusters devueltos pueden ser diferentes. Es importante entonces que el algoritmo sea insensible al orden de los datos, y que el conjunto de clusters devuelto sea siempre el mismo.

- **Alta dimensionalidad.-** Una BD o DW (Data Warehouse) puede contener varias dimensiones o atributos, por lo que es bueno que un algoritmo de clustering pueda trabajar de manera eficiente y correcta no sólo en repositorios con pocos atributos, sino también en repositorios con un alto espacio dimensional, o gran cantidad de atributos.

- **Clustering basados en restricciones.-** Es un gran desafío el agrupar los datos teniendo en cuenta no sólo el comportamiento, sino también que satisfagan ciertas restricciones.

- **Interpretación y uso.-** Los usuarios esperan que los resultados del clustering sean comprensibles, fáciles de interpretar y de utilizar.

Los algoritmos de agrupación o clustering, sólo requieren de la definición previa del vector de características. Una vez establecido dicho vector, los procedimientos de agrupación reciben como datos de entrada los objetos convertidos lógicamente en vectores numéricos a clasificar, de modo que a partir de estos datos de entrada, el algoritmo sin supervisión de ningún tipo y de
forma autónoma, agrupa esos vectores en clases o clusters. Por esta razón también se los denomina “algoritmos de clasificación autoorganizada”.

Los algoritmos de agrupación varían entre sí por el mayor o menor grado de reglas heurísticas que utilizan e, inversamente, por el nivel de procedimientos formales involucrados. Todos ellos se basan en el empleo sistemático de las distancias entre los vectores (objetos a agrupar) así como entre los clusters o grupos que se van haciendo y deshaciendo a lo largo del proceso correspondiente.

2.2.2 Métodos de Clustering

Existen distintos métodos de clustering, tales como:

1. Métodos de Particionamiento
2. Métodos Jerárquicos
3. Métodos basados en la Densidad
4. Métodos basados en Modelos
5. Métodos basados en Grillas

2.2.2.1. Métodos de Particionamiento

Dada una BD de n objetos, un método de particionamiento construye k particiones de datos, en donde cada partición representa un cluster y k <= n. Es decir, se clasifican los datos en k grupos, los cuales satisfacen los siguientes requerimientos:

- ❖ Cada grupo debe contener al menos un objeto.
- ❖ Cada objeto debe pertenecer a un grupo.

Dado k, el número de particiones a construir, un método de particionamiento se crea una partición inicial, y luego se aplica alguna técnica iterativa de reubicación de elementos que intenta mejorar el

particionamiento moviendo objetos de un grupo a otro. El criterio general para un buen particionamiento es el de agrupar en el mismo cluster a objetos cercanos o relacionados, mientras que los correspondientes a distintos grupos serán los considerados alejados o distintos.

Los cluster se forman de manera eficiente aplicando un criterio objetivo de particionamiento conocido como función de “similitud” en donde los objetos en el mismo cluster se consideran similares, y los pertenecientes a distintos clusters se consideran disímiles en relación a sus atributos de la Base de Datos.

Un criterio de agrupación habitualmente utilizado es el de la distancia Euclídea entre vectores, esto es:

$$dE(X_i, X_j) = \sqrt{\sum_{K=1}^{N} (X_{i[K]} - X_{j[K]})^2}$$

Es un ejemplo de método de particionamiento el algoritmo de las K-medias

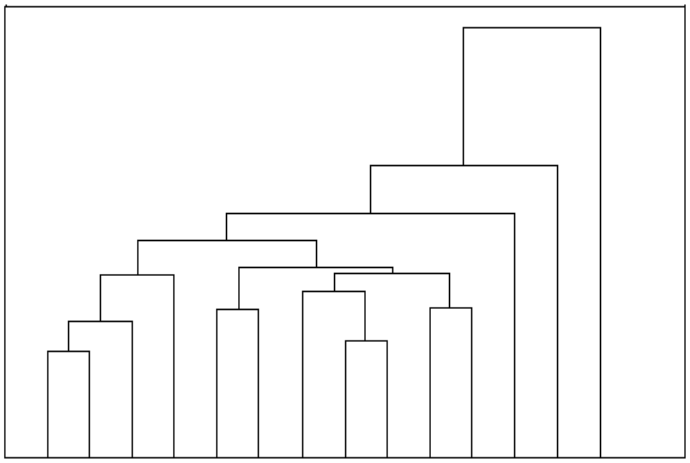
2.2.2.2. Métodos Jerárquicos

Es aquel que crea una descomposición jerárquica del conjunto de objetos dados. Un método de este tipo puede ser considerado como aglomerativo o dispersativo, basándose en cómo se realice la descomposición jerárquica. El criterio aglomerativo, también conocido como bottom-up, comienza generando un grupo para cada uno de los objetos de la base de datos (BD). Luego los grupos se van mezclando y uniendo de manera tal que se encierren en un mismo grupo a aquellos elementos cercanos. Esta operación se sigue hasta llegar a cumplirse determinada condición (se llega al tope de la jerarquía).

El criterio dispersativo o de disgregación, también conocido como top-down, comienza con todos los objetos en un mismo grupo, y en cada una de las sucesivas iteraciones, los objetos se van moviendo a otros grupos.

A continuación presentamos un modelo gráfico representativo del criterio dispersativo o disgregación.

Criterio dispersativo o disgregación



La desventaja que tienen estos métodos, es que una vez realizada una mezcla de grupos, o una vez eliminado un objeto de un grupo, el proceso no puede deshacerse. Esta rigidez es ventajosa en relación al tiempo computacional pero a un costo de no analizar todas las combinaciones posibles de decisiones o elecciones a efectuar.

Existen dos técnicas para mejorar la calidad del clustering jerárquico:

- Efectuar un análisis cuidadoso de los vínculos en cada particionamiento jerárquico o,

- ❖ Integrar aglomeración jerárquica y reubicación iterativa usando en primer lugar un algoritmo de aglomeración jerárquica, y luego refinando el resultado utilizando reubicación iterativa

Algunos ejemplos de este método son el AGNES, DIANA (divisa Análisis), CURE (Clustering using representatives), BIRCH (Balanced Iterative reducing and clustering using hierarchical).

2.2.2.3. Métodos basados en la Densidad o distancias

La mayoría de los métodos de particionamiento se basan en la distancia entre los objetos. Tales métodos pueden encontrar sólo cluster de forma esférica y tener problemas al querer encontrar clusters de formas arbitrarias.

Otros métodos de clustering se basan en el concepto de densidad. La idea es que un cluster dado, continúe creciendo hasta tanto la densidad (número de objetos o puntos de datos) en la vecindad exceda algún umbral; esto es que por cada punto dentro de un cluster dado, la vecindad de un radio dado tiene que contener al menos una cantidad mínima de puntos. Este tipo de método puede ser utilizado

para filtrar datos ruidosos o externos y generar clusters de forma arbitraria.

Dentro de los ejemplos que podemos citar se encuentran DBSCAN (Density Based Spatial Clustering Aplication), OPTICS, DENCLUE.

2.2.2.4. Métodos basados en Modelos

Estos métodos establecen un modelo hipotético para cada uno de los clusters, y buscan la mayor aproximación de los datos al modelo dado. Un algoritmo de este tipo puede ubicar clusters construyendo una función de densidad que represente la distribución espacial de los puntos de datos. También lleva a una forma de determinar automáticamente la cantidad de clusters basados en estadísticas estándares, teniendo en cuenta los puntos exteriores y los “ruidosos” y por lo tanto estableciendo métodos de clustering robustos.

Los métodos basados en modelos siguen dos técnicas principales: una técnica estadística, o una técnica de red neuronal.

2.2.2.5. Métodos basados en Grillas

Estos métodos representan el espacio a través de un número finito de celdas que forman una estructura de grilla. Todas las operaciones de

clustering se llevan a cabo sobre dicha estructura. La ventaja principal de esta técnica es su escaso tiempo de procesamiento, el cual típicamente es independiente de la cantidad de datos, y dependiente sólo de la cantidad de celdas en cada dimensión del espacio cuantificado.

Ejemplos : STING, Wave Clusters, Clique, etc.

2.2.3 Tipos de Clasificación

2.2.3.1 Aglomerativa / Divisiva

En una clasificación aglomerativa se parte inicialmente de los objetos, que se van progresivamente fusionando para tomar particiones sucesivas; en una clasificación divisiva se parte del conjunto total el cual se subdivide progresivamente (de forma dicotómica en general) hasta alcanzar un grado aceptable de subdivisión.

2.2.3.2 Jerárquica / No Jerárquica

En una clasificación no jerárquica se forman grupos homogéneos sin establecer relaciones entre los grupos; en una clasificación jerárquica

los grupos se van fusionando progresivamente, mientras decrece la homogeneidad entre los grupos, cada vez más amplios, que se van formando. Medir esta homogeneidad mediante un índice (distancias) es una característica de la taxonomía numérica. Una clasificación jerárquica es en general aglomerativa.

2.2.3.3 Monotética / Politética

Una clasificación monotética está basada en un característica única (o en unas pocas), que sea muy relevante. Es divisiva, pues los objetos se clasifican en los que tienen la característica y los que no la tienen. Puede dar lugar a clasificaciones poco adecuadas, dada la dificultad de obtener grupos bastante homogéneos y naturales. Una clasificación politética está basada en un número grande de características (en general), y no exige que todos los elementos de una clase posean todas las características, sino un número suficientemente grande para poder justificar analogías entre miembros de una misma clase.

Este tipo de clasificación es aglomerativo.

Hemos mencionado sólo algunos de los aspectos inherentes de una clasificación. Lo que nos interesa destacar son las propiedades aglomerativa y politética de una clasificación jerárquica. La taxonomía numérica, que se empezó a desarrollar hacia los sesenta del siglo veinte, ha podido ser viable y operativa como consecuencia de las posibilidades que ha ofrecido la informática, sobre todo desde la aparición de ordenadores de alta capacidad y velocidad. Se considera que la obra que más ha influido en el enfoque numérico de la clasificación es el libro “Principles of Numerical Taxonomy”, escrito por Sokal y Sneath (1963), en el que exponen “el estudio teórico de la clasificación incluyendo sus bases, principios, procedimientos y reglas”. Posteriormente, otras obras de taxonomía matemática darían soporte teórico a los métodos jerárquicos de clasificación los cuales están relacionados con las “distancias ultramétricas” y sus propiedades.

En líneas generales, la taxonomía numérica intenta construir clasificaciones “naturales”, basadas en la semejanza de los individuos, que se valora partiendo de una adecuada elección de un coeficiente de similaridad.

2.3 METODO K-MEDIAS

El método k-medias es el más popular entre los métodos de agrupamiento basados en suma de cuadrados. Este método es subóptimo y se basa en minimizar el primer criterio que hemos nombrado previamente (minimización de la diagonal).

La idea del método es sencilla; se parte del número de agrupamientos que buscamos k, y se seleccionan inicialmente k muestras de entre las disponibles. Cada una de esas k muestras es la semilla de un agrupamiento, y se asocia el resto de muestras al agrupamiento definido por estas k iniciales, haciendo este agrupamiento por mínima distancia, y se calcula el valor del criterio descrito previamente. Tras este primer paso, se recalculan las medias de cada uno de estos agrupamientos y se usan para volver a agrupar todas las muestras; se calcula el valor del criterio de minimización y si resulta mayor o igual que en la iteración previa, se detiene el proceso. En caso contrario, se sigue iterando.

Este problema de obtener k agrupamientos admite un enfoque desde el punto de vista combinatorio que ha sido estudiado desde hace tiempo y que es conocido como el problema de las k-medias. El problema de las

k-medias se puede ver como una generalización del problema de hallar la media de un conjunto de muestras. Así, si el problema de hallar la media de un conjunto de muestras consiste en hallar el punto del espacio de representación cuya suma de distancias al conjunto de muestras sea mínima, el problema de las k-medias trata de hallar los k puntos (representantes) del espacio de representación tal que la suma de distancias de cada muestra a su representante más cercano sea mínima.

Así, si P es el conjunto de muestras sobre un cierto espacio de representación E con una cierta distancia d definida, dicha formulación viene expresada por hallar el conjunto de puntos $Q \subset E$, con $|Q| = k$, que cumpla la ecuación

$$Q = \underset{Q \subset \mathbf{E}, |Q|=k}{\operatorname{argmin}} \sum_{p \in P} \min_{q \in Q} d(p, q)$$

Evidentemente, la solución de este problema es una solución al problema de obtener k agrupamientos, ya que cada representante generará un agrupamiento formado por las muestras más cercanas al mismo.

2.3.1 Algoritmo K-Medias

El nombre de este algoritmo hace referencia a que existen K clases o patrones, siendo necesario por lo tanto, conocer a priori el número de clases existentes. Es un algoritmo sencillo, pero muy eficiente, siempre que el número de clases se conozca a priori con exactitud. Por su sencillez y robustez, es muy utilizado.

Partiendo de un conjunto de objetos a clasificar X1, X2, ... , XP (habrá un vector de características Xi por cada observación a clasificar), el algoritmo de las K-medias realiza las siguientes operaciones :

1.- Establecido previamente el número exacto de clases existentes, digamos K, se escogen al azar entre los elementos a agrupar K vectores, de forma que se van a construir los centroides (al ser éstos los únicos elementos) de las K clases.

Es decir :

$$\alpha 1 : Z1(1) ; \alpha 2 : Z2(1) , \ldots , \alpha k : Zk(1)$$

en donde se ha indicado entre paréntesis el índice iterativo o número de iteración de este algoritmo.

2.- Como se trata de un proceso recursivo con un contador n, en la iteración genérica n se distribuyen todas las muestras { X } 1<= j <= P entre las K clases, de acuerdo con la siguiente regla :

$$X \in \alpha j(n) \text{ sii } \| X - Zj(n) \| < \| X - Zi(n) \| \quad \forall\, i = 1,2,\ldots,K \quad \{ i \neq j \}$$

en donde se han indexado las clases (que son dinámicas) y sus correspondientes centroides.

3.- Una vez redistribuídos los elementos a agrupar entre la diferentes clases, es necesario recalcular o actualizar los centroides de las clases.

El objetivo en el cálculo de los nuevos centroides es minimizar el índice de rendimiento siguiente:

$$Ji = \sum_{X \in \alpha i(n)} \| X - Zi(n) \|^2 \quad \text{con } i=1,2,\ldots,K$$

Este índice se minimiza utilizando la media muestral o aritmética de α i (n):

$$Zi(n+1) = \frac{1}{Ni(n)} \sum_{X \in \alpha i(n)} X \quad ; \quad \text{con } i=1,2,\ldots,K$$

Siendo Ni(n) el número de elementos de la clase ai en la n-ésima iteración.

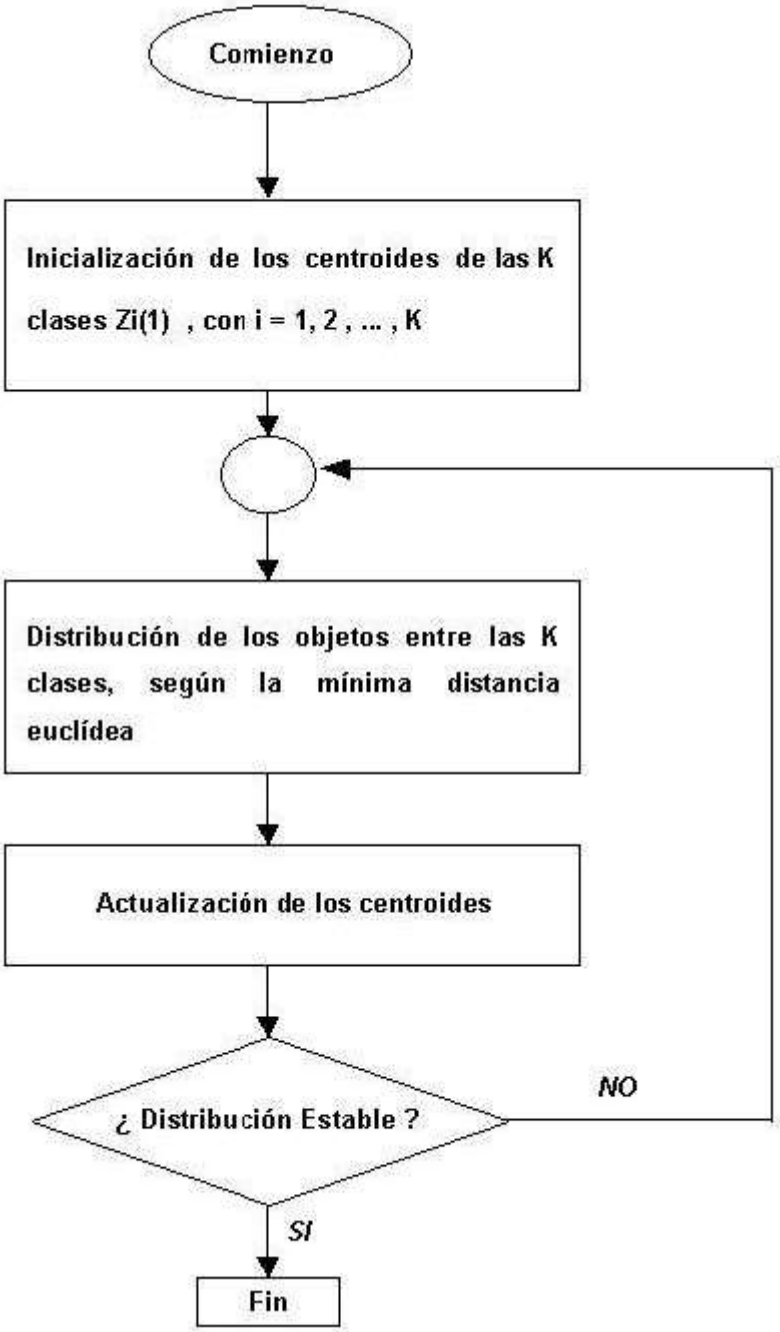
4.- Se comprueba si el algoritmo ha alcanzado una posición estable. Es decir, si se cumple:

$$Zi(n+1) = Zi(n) \quad \forall \; i = 1, 2, \ldots, K$$

Si se cumple, el algoritmo finaliza. En el caso contrario, vuelve al paso 2.

El algoritmo de las K-Medias es simple y muy eficiente si el número de clases se conoce a priori con exactitud. Es decir, es muy sensible al parámtero K. Un valor de K superior al número real de clases dará lugar a clases ficticias, mientras que un K inferior producirá menos clases de las reales.

2.3.2 Representación Gráfica del algoritmo K-Medias



2.3.3 MEDIDAS DE DISIMILITUD ENTRE CADENAS

Como hemos visto en la Sección 1.1, para la aplicación de técnicas no paramétricas de clasificación es necesaria la definición de una medida de disimilitud entre los objetos. Dicha medida, debe cumplir una serie de condiciones para garantizar la corrección de los resultados de clasificación. Así, dados dos objetos codificados cualesquiera o1 y o2, la medida de disimilitud d debe cumplir las siguientes condiciones:

- $d(o1, o2) \geq 0$
- $d(o1, o1) = 0$
- $d(o1, o2) = d(o2, o1)$

Si además de estas tres condiciones se cumple la condición de desigualdad triangular:

$$\mathbf{d}(o_1, o_2) + \mathbf{d}(o_2, o_3) \geq \mathbf{d}(o_1, o_3) \qquad \forall\ \mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3$$

entonces se tiene que d es una métrica, y se habla de ella como una medida de distancia.

Cuando los objetos se representan mediante vectores, existen gran cantidad de distancias. En el caso de la representación por cadenas

también ha habido diversas propuestas de medidas de disimilitud entre cadenas, algunas de las cuales vamos a detallar a continuación.

2.4 DISTANCIA DE EDICIÓN

El problema de hallar la disimilitud entre dos cadenas de caracteres sobre un determinado alfabeto puede verse como una cantidad numérica que determine el “esfuerzo" necesario para convertir una cadena en otra. Así, cuanto menor sea el esfuerzo necesario para esa transformación, más próximas estarán estas cadenas. A la hora de definir como se realiza dicha transformación, se pueden definir multitud de operadores básicos que la realicen. Históricamente, se han definido como operadores básicos los de sustitución, inserción y borrado. Dadas las cadenas s y t a comparar, de longitudes $|s|$ y $|t|$, $s = s_1 s_2, \ldots s_{|s|}$ y $t = t_1, t_2, \ldots, t_{|t|}$, el operador de sustitución empareja un símbolo de s con un símbolo de t. El operador de inserción hace que un símbolo en la cadena s no este emparejado con ningún símbolo de t. El operador de borrado hace que un símbolo de t no se empareje con ninguno de s (es, por tanto, el dual de inserción).

En la Figura 1.2 podemos ver la representación gráfica habitual de las operaciones de edición, donde λ representa la cadena vacía (de manera que (λ; b) representa la inserción y (a;λ) representa el borrado).

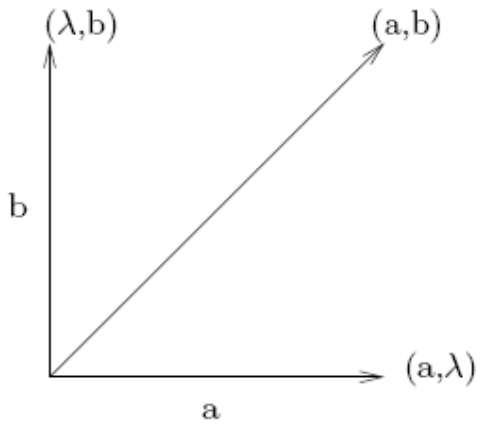


Figura 1.2

2.5 DISTANCIA DE LEVENSHTEIN

Sea F una función arbitraria de costo que asigne a cada operación de edición (ϕ , Ω) un número real no negativo, F(ϕ , Ω). Se puede extender F a la secuencia S definiendo:

$$F(S) = \sum_{j=1}^{n} F(S_j) \text{ si } n \geq 1 \qquad \text{y} \qquad F(S) = 0 \text{ si } n = 0$$

Se denomina distancia de Levenshtein, DL(S,T), entre las cadenas S y T, a todas las secuencias de edición que transformen X en Y.

En este trabajo se ha seguido el criterio de que el costo de cualquier operación de edición se considera como unitario.

La mínima distancia de edición de dos cadenas s y t se define como el mínimo número de mutaciones requeridas para cambiar s a t, donde una mutación puede ser uno de los siguientes cambios:

- ✓ Cambiar una letra por otra
- ✓ Insertar una letra
- ✓ Eliminar una letra

2.5.1 Implementación del Algoritmo de Levenshtein.

Pasos	Detalles
1	Establecer una variable **n** que contenga el tamaño de la cadena de caracteres inicial **s.** Establecer una variable **m** que contenga el tamaño de la cadena de caracteres final **t.** Si **n** = 0, devuelva el valor de **m** y salga. Si **m** = 0, devuelva el valor de **n** y salga. Se construye una matriz **d** de **m** filas x **n** columnas.
2	Inicializar la primera fila de 0…..**n** Inicializar la primera columna de 0…..**m**
3	Examine cada caracter de **s** (i de 1 a **n**).
4	Examine cada caracter de **t** (j de 1 a **m**).
5	Si **s [i]** es igual a **t [j]**, el costo es 0. Si **s [i]** no es igual a **t [j]**, el costo es 1.
6	Establecer la celda **d [i , j]** de la matriz igual al mínimo de: a. La celda inmediatamente sobre 1: **d [i-1, j] + 1.** b. La celda inmediatamente a la izquierda mas 1: **d [i, j-1] + 1.** c. La celda diagonal sobre la izquierda más el costo: **d[i-1, j-1] + cost.**
7	Después de que los pasos de iteración [3,4,5,6] son completados, la **distancia** es encontrada en la celda **[n,m]**

Ejemplo:

Este ejemplo muestra como trabaja la distancia de Levenshtein, cuando la cadena de caracteres fuente es “GUMBO” y la cadena de caracteres destino es “GAMBOL”.

Pasos 1 y 2

		G	**U**	**M**	**B**	**O**
	0	1	2	3	4	5
G	1					
A	2					
M	3					
B	4					
O	5					
L	6					

Pasos 3 al 6, cuando i = 1

		G	**U**	**M**	**B**	**O**
	0	1	2	3	4	5
G	1	0				
A	2	1				
M	3	2				
B	4	3				
O	5	4				
L	6	5				

Pasos 3 al 6, cuando i = 2

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1			
A	2	1	1			
M	3	2	2			
B	4	3	3			
O	5	4	4			
L	6	5	5			

Pasos 3 al 6, cuando i = 3

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2		
A	2	1	1	2		
M	3	2	2	1		
B	4	3	3	2		
O	5	4	4	3		
L	6	5	5	4		

Pasos 3 al 6, cuando i = 4

		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	
A	2	1	1	2	3	
M	3	2	2	1	2	
B	4	3	3	2	1	
O	5	4	4	3	2	
L	6	5	5	4	3	

Pasos 3 al 6, cuando i = 5

		G	**U**	**M**	**B**	**O**
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	2	2	1	2	3
B	4	3	3	2	1	2
O	5	4	4	3	2	1
L	6	5	5	4	3	**2**

Distancia

Paso 7

La distancia está en la parte inferior derecha en la esquina de la matriz, en este caso es 2. Esto corresponde a nuestra realización intuitiva de que la palabra “GUMBO” se puede transformar en “GAMBOL”, por una sustitución de una “A “ por una “ U “ y agregar una “ L “ (una sustitución y una inserción, igual a dos cambios).

Capítulo 3

3.- Agrupación de las descripciones

Debido a que en el capítulo anterior revisamos los métodos de particionamiento y agrupación de los datos esta parte revisaremos los conceptos de muestreo, los mismos que nos servirán como herramienta para la obtención y procesamiento de los datos. Como ya veremos más adelante el muestreo será utilizado para determinar sobre una gran cantidad de datos una proporción que sea representativa para el análisis y en base a los resultados tomar decisiones sobre el total de los datos.

Así también, veremos la terminología, los tipos de muestreo, fórmulas de cálculo, de donde obtuvimos los datos, el análisis de los mismos y el criterio de discriminación para el agrupamiento de los registros.

3.1 Muestreo

Concepto.-

El muestreo es una herramienta de la investigación científica. Su función básica es determinar que parte de una realidad en estudio (población o universo) debe examinarse con la finalidad de hacer inferencias sobre dicha población. El error que se comete debido a hecho de que se obtienen conclusiones sobre cierta realidad a partir de la observación de solo una parte de ella, se denomina error de muestreo. Obtener una muestra adecuada significa lograr una versión simplificada de la población, que reproduzca de algún modo sus rasgos básicos.

3.1.1 Terminología

Población objetivo: Conjunto de individuos de los que se quiere obtener una información.

Unidades de muestreo: número de elementos de la población, que se van a estudiar. Todo miembro de la población pertenecerá a una y sólo una unidad de muestreo.

Unidades de análisis: Objeto o individuo del que hay que obtener la información.

Marco muestral: lista de unidades o elementos de muestreo.

Muestra: conjunto de unidades o elementos de análisis sacados del marco.

3.1.2 Muestreo probabilístico

El método otorga una probabilidad conocida de integrar la muestra a cada elemento de la población, y dicha probabilidad no es nula para ningún elemento.

Los tipos de muestreo probabilística son:

- Muestreo Aleatorio Simple
- Muestro Sistemático
- Muestreo Aleatorio Estratificado
- Muestreo por conglomerados (clusters)

3.1.3 Muestreo aleatorio simple

Muestreo equiprobabilístico: Si se selecciona una muestra de tamaño *n,* de una población de *N* unidades, cada elemento tiene una probabilidad de inclusión igual y conocida de *n/N.*

Ventajas:

- Sencillo y de fácil comprensión.
- Cálculo rápido de medias y varianzas.
- Se basa en la teoría estadística, y por tanto existen paquetes informáticos para analizar los datos.

Desventajas:

- Requiere que se posea de antemano un listado completo de toda la población.
- Cuando se trabaja con muestras pequeñas es posible que no represente a la población y conocida de *n/N*.

3.1.4 Muestreo sistemático

Procedimiento:

- Conseguir un listado de N elementos.
- Determinar un tamaño de muestra n.
- Definir un intervalo de salto k, K=N/n.
- Elegir un número aleatorio, r, entre 1 y k (r = arranque aleatorio).

- Seleccionar los elementos de la lista.

Ventajas:

- Fácil de aplicar.
- No siempre es necesario tener un listado de toda la población.
- Cuando la población está ordenada siguiendo una tendencia conocida, asegura una cobertura de unidades de todos los tipos.

Desventajas:

- Si la constante de muestreo está asociada con el fenómeno de interés, se pueden hallar estimaciones sesgadas.

3.1.5 Muestreo aleatorio estratificado

El azar no es una garantía de representatividad. Este muestreo pretende asegurar la representación de cada grupo en la muestra. Cuanto más homogéneos sean los estratos, más precisas resultarán las estimaciones.

Ventajas:

- Tiende a asegurar que la muestra represente adecuadamente a la población en función de unas variables seleccionadas.

- Se obtienen estimaciones más precisas.

Desventajas:

- Se ha de conocer la distribución en la población de las variables utilizadas para la estratificación.
- Los análisis son complicados, en muchos casos la muestra tiene que ponderarse (asignar pesos a cada elemento).

3.1.6 Muestreo por conglomerados

Los conglomerados se caracterizan porque la variación en cada grupo es menor que la variación entre grupos.
La necesidad de listados de las unidades de una etapa se limita a aquellas unidades de muestreo seleccionadas en la etapa anterior.

Ventajas:

- Es muy eficiente cuando la población es muy grande y dispersa. Reduce costes.
- No es preciso tener un listado de toda la población, sólo de las unidades primarias de muestreo.

Desventajas:

- El error estándar es mayor que en el muestro aleatorio simple o estratificado.
- El cálculo del error estándar es complejo.

3.2 Cálculo del tamaño muestral

Cada estudio tiene un tamaño muestral idóneo, que permite comprobar lo que se pretende con una seguridad aceptable y el mínimo esfuerzo posible. Para el cálculo del tamaño muestral en cada tipo de estudio existe una fórmula estadística apropiada. Se basan en el error estándar, que mide el intervalo de confianza de cada parámetro que se analiza (media aritmética, porcentaje, diferencia de medias, etc.). La precisión estadística aumenta (el error estándar disminuye) cuando el tamaño muestral crece.

Se trata de un caso especial debido a que solo tenemos una cantidad N de cadenas de caracteres y hemos considerado que nuestra población de datos es totalmente diferente, por lo tanto se trabajó con el muestreo aleatorio simple. La fórmula para calcular el tamaño de muestra, está basada en un error de muestreo dado y un coeficiente de confianza para

proporciones, ya que necesitamos estimar una proporción que sea representativa en relación al total de nuestra población. Ambos valores fueron tomados con el fin de maximizar el tamaño de la muestra (error = 0.03 y coeficiente de confianza = 0.95) y ser más precisos en el análisis. Debido a que queremos maximizar el tamaño de la muestra tomamos a ***p*** y ***q*** con un valor de 0.5 ya que ese es el valor que al reemplazarlo en la fórmula nos da como resultado el tamaño más grande del tamaño de muestra.

3.2.1 Fórmula para el cálculo del tamaño de muestra.

A continuación mostramos la fórmula que se utilizó para el cálculo del tamaño de muestra, en la cual detallamos el significado de cada una de las variables.

$$n = \frac{\lambda_\alpha^2 NPQ}{(N-1)e_\alpha^2 + \lambda_\alpha^2 PQ}$$

n = tamaño de la muestra.

λ_α^2 = Coeficiente de confianza. (λ_α^2 = 0.95).

N = tamaño de la población, (total de los datos).

e_α^2 = error de muestreo. (e_α^2 =0.03).

Entonces, para poblaciones grandes o fracción de muestreo pequeña ($N \to \infty$), el valor máximo de n se obtiene para **P=Q=1/2**. Luego, cuando se estiman proporciones y no se conoce el valor de la proporción poblacional P ni se tiene una aproximación suya (proporcionada por una encuesta similar, por una encuesta piloto, por la misma encuesta realizada anteriormente ni por ningún otro método), puede tomarse **P=1/2**, siempre y cuando no se obtenga un tamaño muestral n demasiado grande en términos de coste.

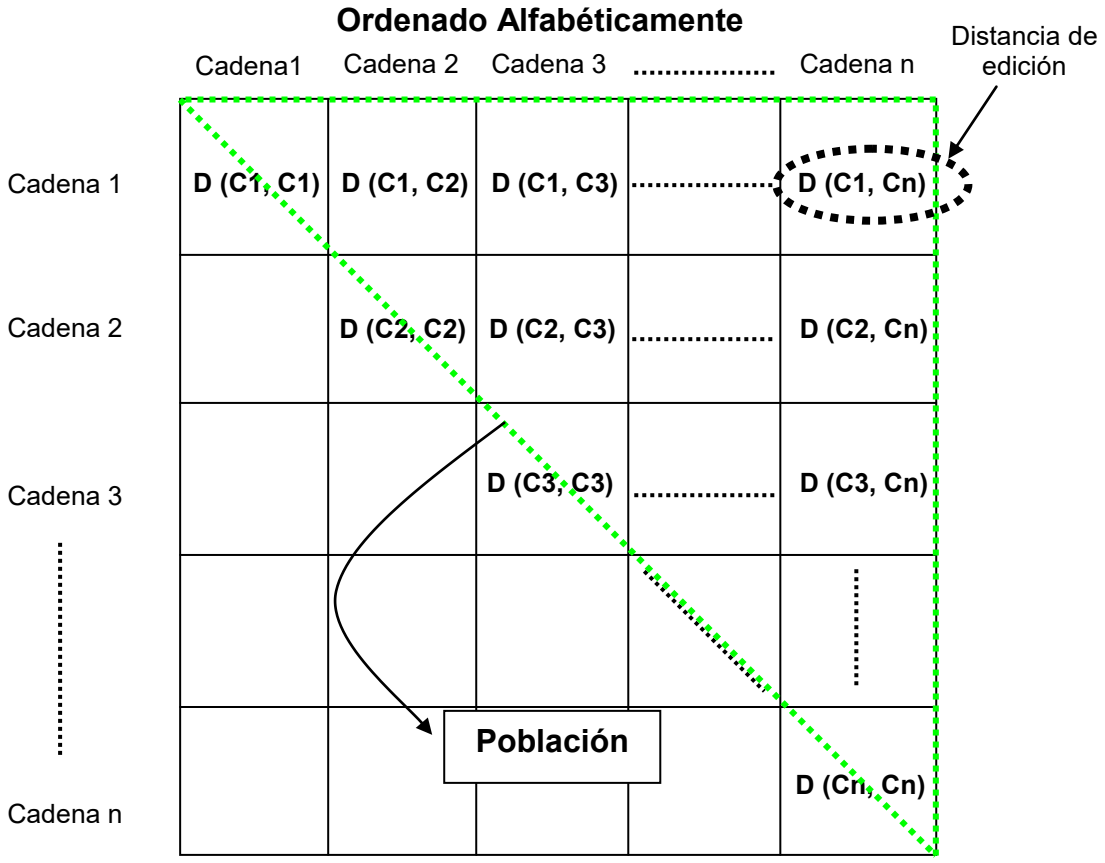
3.2.2 Obtención de los datos

Los datos que se tomaron para el desarrollo de esta tesis se asemeja mucho a lo que pueda existir en un servicio de Aduanas. Por lo tanto la mejor fuente considerada para la obtención de los mismos, fueron los listados de artículos que varias compañías ofrecen a través del Internet.

Se elaboro una lista con 1858 artículos de diferentes categorías de productos. Cabe recalcar que de este listado como en cualquier otro pueden existir datos repetidos. Por lo tanto, esto fue considerado en el desarrollo de la aplicación informática.

3.2.3 Análisis de los datos

Luego de haber calculado el tamaño de muestra, utilizamos la distancia de levenshtein como herramienta, para obtener la mínima distancia de conversión (inserción, borrado o reemplazo) entre dos cadenas de caracteres. Para analizar los datos tomamos las cadenas de caracteres y las ubicamos en dos vectores, A y B. Entonces, realizamos una comparación secuencial para obtener las distancias que existen entre cada una de las cadenas para luego poder aplicar el criterio de agrupación en base al parámetro obtenido. El proceso de comparación secuencial se resume en una matriz en la cual sólo consideramos la triangular superior para evitar repeticiones y por lo tanto resultados erróneos, de la siguiente forma:



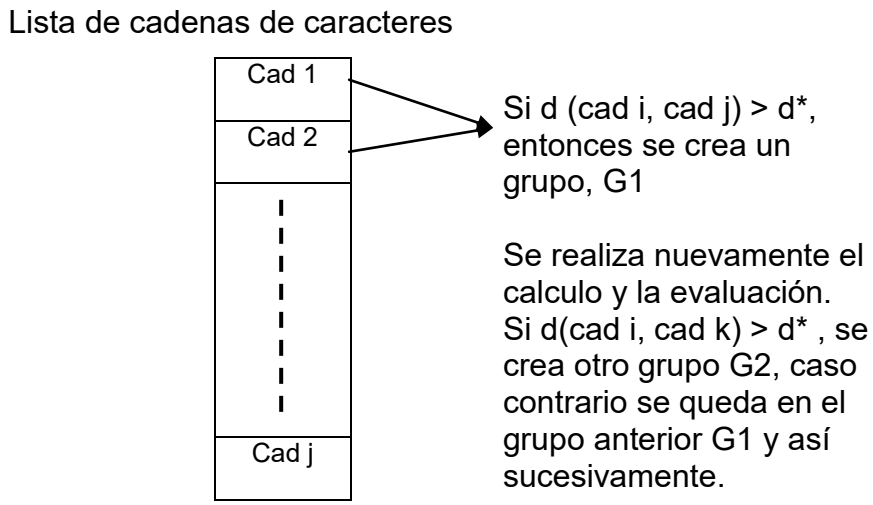
Como se observa en la figura anterior, los cuadros denotados con una trama de puntos son las coordenadas de la matriz que consideramos para el análisis y cálculo de distancias secuenciales evitando así las repeticiones de los resultados.

3.3 Criterio de agrupación

Dado una lista de cadenas de caracteres ordenadas alfabéticamente para realizar la agrupación se utilizara la siguiente regla:

Si	d <= d*	Los pares son iguales. (1)
Si	d > d*	Los pares no son iguales,

En donde d es la distancia entre las cadenas de caracteres, y d* es la distancia óptima. Llamamos distancia óptima a aquella distancia que nos genere el menor error de discriminación.



El error de discriminación de la regla (1) está dado por la siguiente fórmula:

Error de discriminación si d* = j

$$Error(d^* = j) = \sum_{i=1}^{j} 1 - \Pr ob(simi \setminus d) + \sum_{i=j+1}^{n} \Pr ob(simi \setminus d)$$

En donde $\Pr ob(simi \setminus d)$ es la probabilidad de que dos cadenas de caracteres sean similares a una determinada distancia y 1- $\Pr ob(simi \setminus d)$ es el complemento (lo contrario).

Error(d*=j) = Valor Esperado de Pares que no son similares y la distancia entre ellos es <= d* + Valor Esperado de Pares que son iguales y la distancia entre ellos es > d*

	D(C1,C2) <= d*	**D(C1,C2) > d***
C1 y C2 iguales	0	$\sum_{i=j+1}^{n} \Pr ob(simi / d)$
C1 y C2 no iguales	$\sum_{i=1}^{j} 1 - \Pr ob(simi / d)$	0

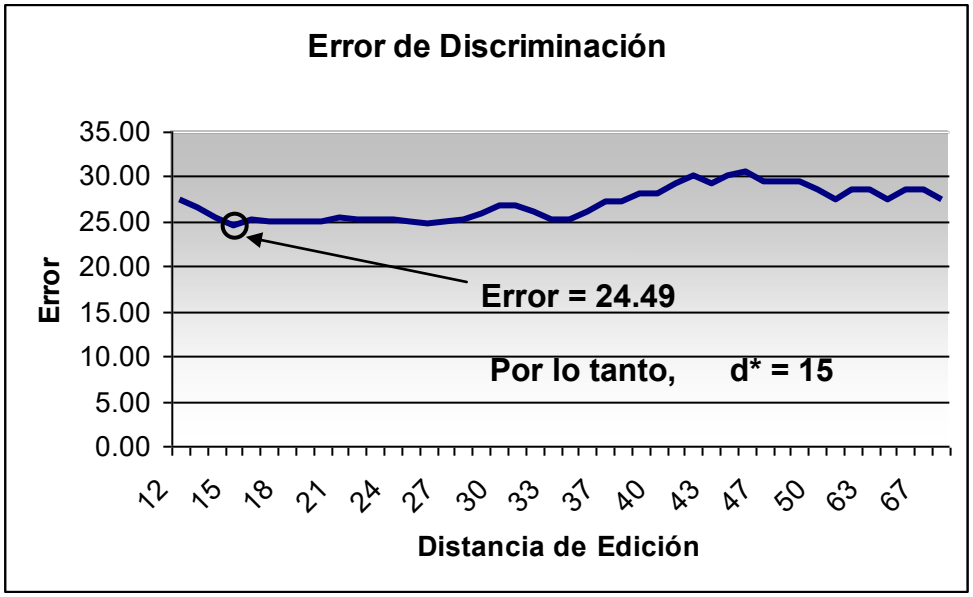
Por lo tanto, la distancia óptima esta dada por:

d* = **min** (error de discriminación)

A continuación presentamos los resultados del análisis manual respecto a las similaridades entre las cadenas de caracteres de la muestra (238 datos).

dist	Simil	Difer	Prob(simi/d)	1-Prob(Simi/d)	Probabilidad	Error
12	1	0	1	0	0.00840336	27.49
13	1	0	1	0	0.00840336	26.49
14	1	0	1	0	0.00840336	25.49
15	**1**	**0**	**1**	**0**	**0.00840336**	**24.49**
16	1	4	0.2	0.800000028	0.00840336	25.09
17	6	5	0.545454545	0.45454545	0.05042017	25.00
18	7	6	0.538461538	0.461538467	0.05882353	24.92
19	6	8	0.428571428	0.571428569	0.05042017	25.06
20	3	2	0.6	0.39999999	0.02521008	24.86
21	3	13	0.1875	0.812499998	0.02521008	25.49
22	7	3	0.7	0.300000004	0.05882353	25.09
23	7	7	0.5	0.500000008	0.05882353	25.09
24	7	8	0.466666666	0.533333343	0.05882353	25.16
25	12	7	0.631578947	0.368421047	0.10084034	24.89
26	12	8	0.6	0.399999997	0.10084034	24.69
27	3	5	0.375	0.624999996	0.02521008	24.94
28	2	4	0.333333333	0.666666679	0.01680672	25.28
29	1	3	0.25	0.750000029	0.00840336	25.78
30	0	5	0	1	0	26.78
31	3	3	0.5	0.5	0.02521008	26.78
32	4	1	0.8	0.200000019	0.03361345	26.18
33	3	0	1	0	0.02521008	25.18
34	2	2	0.5	0.5	0.01680672	25.18
36	0	1	0	1	0	26.18
37	0	1	0	1	0	27.18
38	1	1	0.5	0.500000059	0.00840336	27.18
39	0	2	0	1	0	28.18
40	2	2	0.5	0.5	0.01680672	28.18
41	0	3	0	1	0	29.18
42	0	1	0	1	0	30.18
43	1	0	1	0	0.00840336	29.18
45	0	1	0	1	0	30.18
46	1	2	0.333333333	0.666666719	0.00840336	30.51
47	3	0	1	0	0.02521008	29.51
48	1	1	0.5	0.500000059	0.00840336	29.51
49	1	1	0.5	0.500000059	0.00840336	29.51
50	1	0	1	0	0.00840336	28.51

51	1	0	1	0	0.00840336	27.51
62	0	2	0	1	0	28.51
63	1	1	0.5	0.500000059	0.00840336	28.51
64	1	0	1	0	0.00840336	27.51
65	0	1	0	1	0	28.51
67	1	1	0.5	0.500000059	0.00840336	28.51
68	2	0	1	0	0.01680672	27.51
69	0	1	0	1	0	28.51
70	0	1	0	1	0	29.51
72	2	0	1	0	0.01680672	28.51
73	3	0	1	0	0.02521008	27.51
74	1	0	1	0	0.00840336	26.51
75	1	0	1	0	0.00840336	25.51
78	0	1	0	1	0	26.51
79	0	1	0	1	0	27.51
80	1	0	1	0	0.00840336	26.51
81	1	0	1	0	0.00840336	25.51

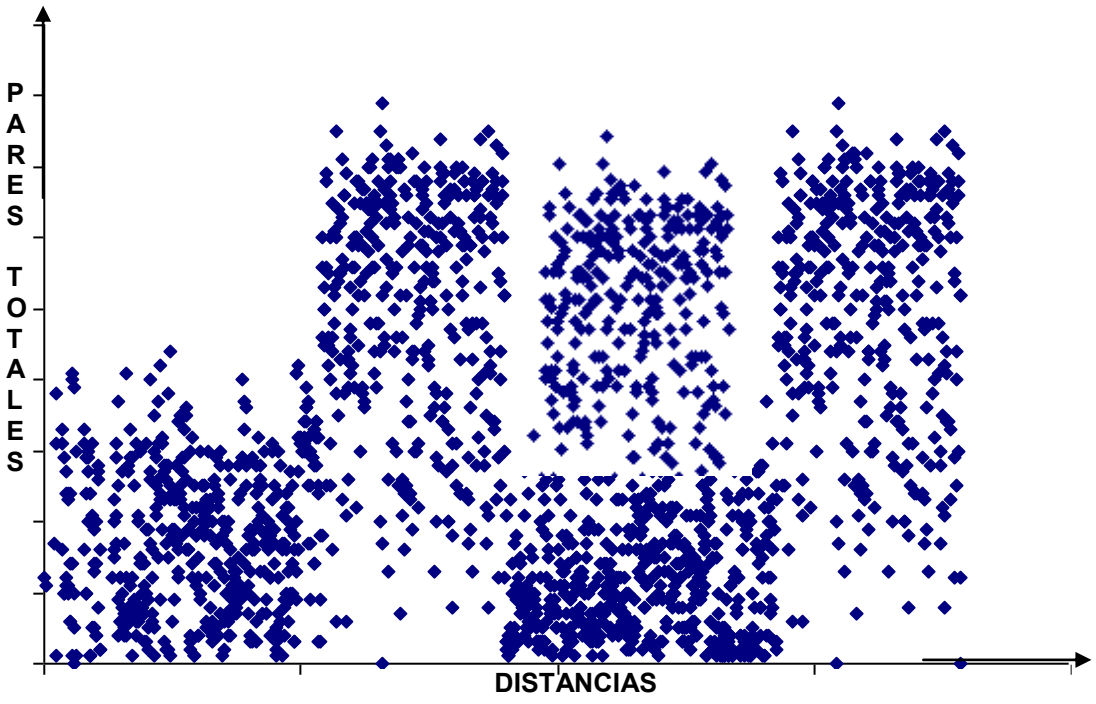


Como podemos observar en la tabla mostrada anteriormente encontramos una fila sombreada, la cual nos indica que genera el menor error de nuestro análisis y por lo tanto la distancia óptima (d*). Si

observamos el gráfico encontramos que efectivamente la menor distancia de edición es 15 debido a que fue generada por el menor error de discriminación (24.49).

Aplicando esto en nuestra población total, porque recordemos que el análisis sobre la muestra es con el objetivo de proyectarlo sobre una cantidad de datos mucho mayor, en este caso nuestra población; nos da como resultado 202 grupos categóricos, en los cuales se encuentra distribuida la población total.

Para explicarlo gráficamente queremos llegar de una gran cantidad de registros, a una cantidad menor de grupos que representen dichos registros. Es decir, desde esto:



CAPITULO IV

4. Introducción

En esta parte observaremos parte del codigo fuente para la agrupación automática de descripciones de productos, una muestra de las descripciones de los productos, grupos que se forman, un breve comentario de lo que es el record linkage (otra técnica para enlazar datos), y tambien las conclusiones y recomendaciones para este trabajo.

4.1 Record Linkage

El record linkage en una definición clara y concisa, es una fusión rápida y eficiente de múltiples ficheros con errores y con falta de datos.

Si partimos de un problema dado como identificar las distintas apariciones de coincidencias en una entidad en ficheros de registros y:

- Registros no identificados de manera global,
- Error en los datos.

La solución para nuestro problema es utilizar el record linkage. Las ventajas que este nos presenta es que no depende de un identificador global y es tolerante a errores en el registro e incompletitudes en los datos. La desventaja es que es un método estadístico e inexacto.

Otras de las desventajas que puede presentar el record linkage las detallamos a continuación:

Tratamiento previo de los datos (homogeneidad).

Formato

Errores de introducción.

Función de comparación de registros.

Determina si dos registros son similares.

Modelo probabilística.

Requiere un límite estadístico.

Realizar todas las comparaciones entre registros:

Coste cuadrático.

Por lo general inabordable.

Fases del proceso de record linkage

1.- Métodos de detección de similitudes (acercamiento de registros similares):

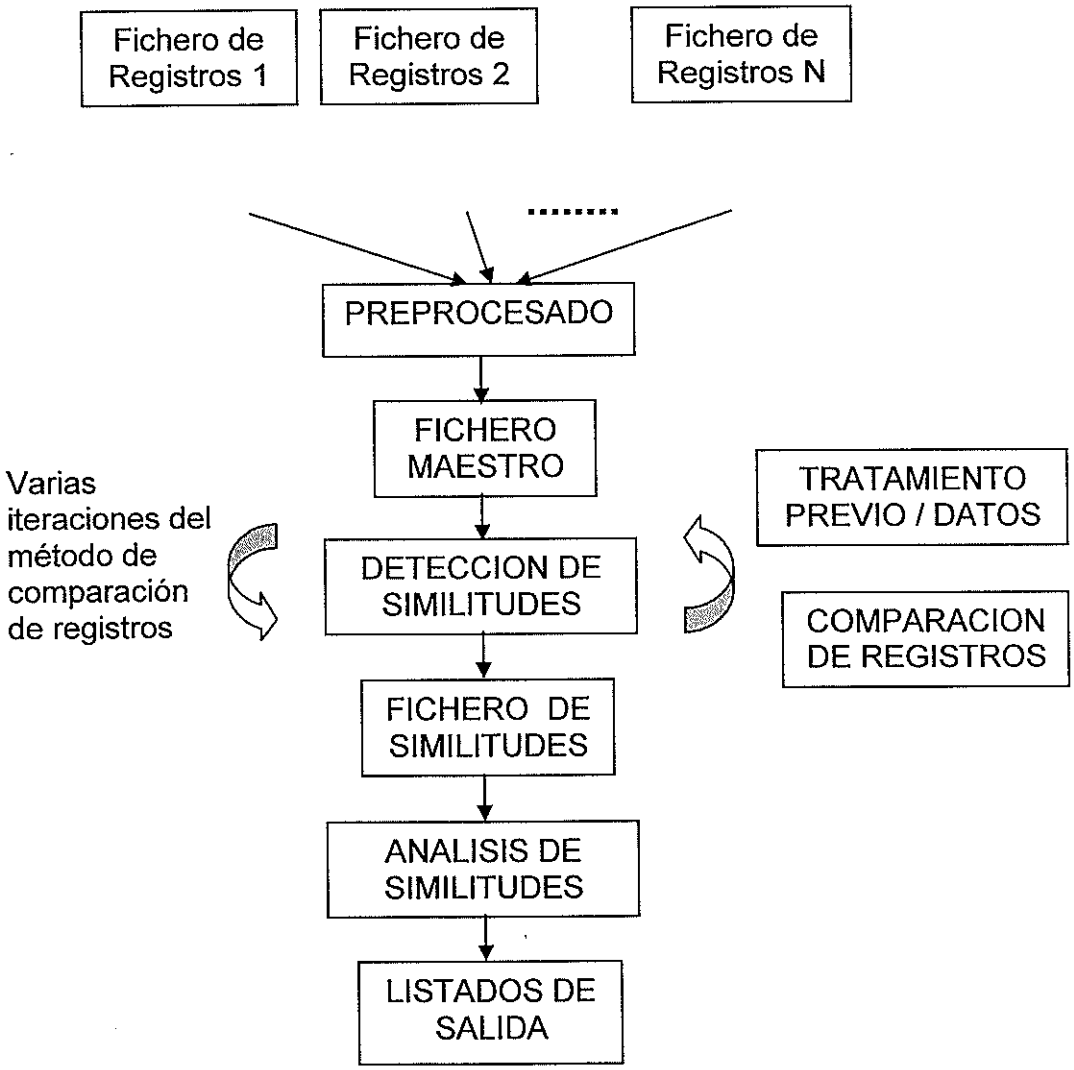
- Blocking
- Window

Esto implica la reducción del coste computacional del proceso y perdida de precisión en la detección de similitudes.

2.- Métodos de Análisis de las relaciones de similitud:

- Clustering
- Implica la inducción o discriminación de similitudes.

Estructura gráfica del proceso



Ref. [9]

Características del Record Linkage

Dentro de las características del Record Linkage encontramos las siguientes:

- Precisión.
- Esfuerzo de revisión.
- Tiempo de ejecución.

La precisión

La perdida precisión para los casos correctos detectados se produce en la detección y en el análisis de similitudes.

Esfuerzo de revisión

El análisis de similitudes tiene un impacto significativo ya que existen casos en los cuales un par de registros pueden ser de la misma entidad y el sistema no los haya detectado y estaríamos perdiendo un integrando más de una determinada entidad.

Tiempo de Ejecución

Dentro del tiempo de ejecución se involucran múltiples ficheros y por lo tanto un número elevado de registros lo que conlleva a días para poder realizar un cruce de información para así verificar la efectividad del método.

Modelo Estándar

En esta sección, describiremos el alcance estándar del record linkage. Consideremos una base de datos de registros los cuales queremos duplicar. Dejemos que cada registro sea representado por un grupo de atributos. Consideremos un par candidato, denotado por y, donde y puede tomar valores desde [-1; 1]. El valor de 1 significa que el registro en el par indicado refiere a la misma entidad o grupo, y un valor de -1 significa que el registro en el par indicado refiere a una entidad diferente. Sea x = (x1; x2;.....; xn) una variable que denota un vector de cuentas de similaridades entre los atributos correspondiendo a los registros en el par candidato. Entonces, en el alcance estándar, la distribución de probabilidad de y dado x, está definido por un modelo de Bayes o regresión logística:

$$f(\mathbf{x}) = \log \frac{P(y=1|\mathbf{x})}{P(y=-1|\mathbf{x})} = \lambda_0 + \sum_{i=1}^{n} \lambda_i x_i$$

f(x) es conocida como la función discriminante. λ i , para $0 \leq i \leq n$, son los parámetros de el modelo. Dados estos parámetros y el atributo de similaridad correspondiente al vector x, un par candidato y puede ser positivo (compatible) si f(x) > 0 y negativo en caso contrario. El descenso de la pendiente es usado para encontrar los parámetros que aumentan al máximo la probabilidad condicional de y dado x, es decir, P(y|x).

Modelo Colectivo

La diferencia básica entre el modelo estándar y el modelo colectivo es que el modelo colectivo no toma decisiones independientemente sobre el par. Mas bien, hace una decisión colectiva para todos los pares candidatos, proyectando información a través de los valores del atributo compartido (x).

Campos aleatorios condicionales

Los campos aleatorios condicionales son modelos gráficos indirectos los cuales definen la probabilidad condicional de un grupo de variables de salida Y dado por un grupo de variables de entrada X. Formalmente,

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \prod_{c \in C} \phi_c(\mathbf{y}_c, \mathbf{x}_c),$$

donde C es el conjunto de grupos en el gráfico, y yc y xc denotan el subgrupo de variables que participan en el grupo c. ϕc , conocido como un grupo potencial, es una función de las variables involucradas en el grupo c. Zx es la normalización constante. Tipicamente, ϕc está definida como una combinación lineal de rasgos sobre c, i.e., ϕc (yc , xc) = exp $\sum$l λ lc f lc (yc , xc), donde f lc conocida como la función de rasgos, es una función de variables involucradas en el grupo c, y λ lc son los pesos de los rasgos.

En algunas áreas, en lugar de tener diferentes parámetros (pesos de los rasgos) para cada grupo en el gráfico, los parámetros de un campo aleatorio condicional son atados por un grupo (patrón) que se repite en el gráfico. Llamamos a cada patrón una plantilla de grupo relacional. Cada grupo c compatible a una plantilla de grupo t es llamado una instancia de la plantilla. La distribución de probabilidad entonces puede ser especificada como:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \sum_{t \in T} \sum_{c \in C_t} \exp \sum_{l} \lambda_{lt} f_{lt}(\mathbf{y}_c, \mathbf{x}_c)$$

donde T es el grupo de todas las plantillas, Ct es el conjunto de grupos los cuales satisfacen la plantilla t, y f tc λ tc son respectivamente la función de rasgos y los pesos de los rasgos perteneciendo a la plantilla t. Porque el parámetro, la función de los rasgos y los parámetros, varían sobre la plantilla de los grupos y no por grupos individuales.

4.2 CONCLUSIONES Y RECOMENDACIONES

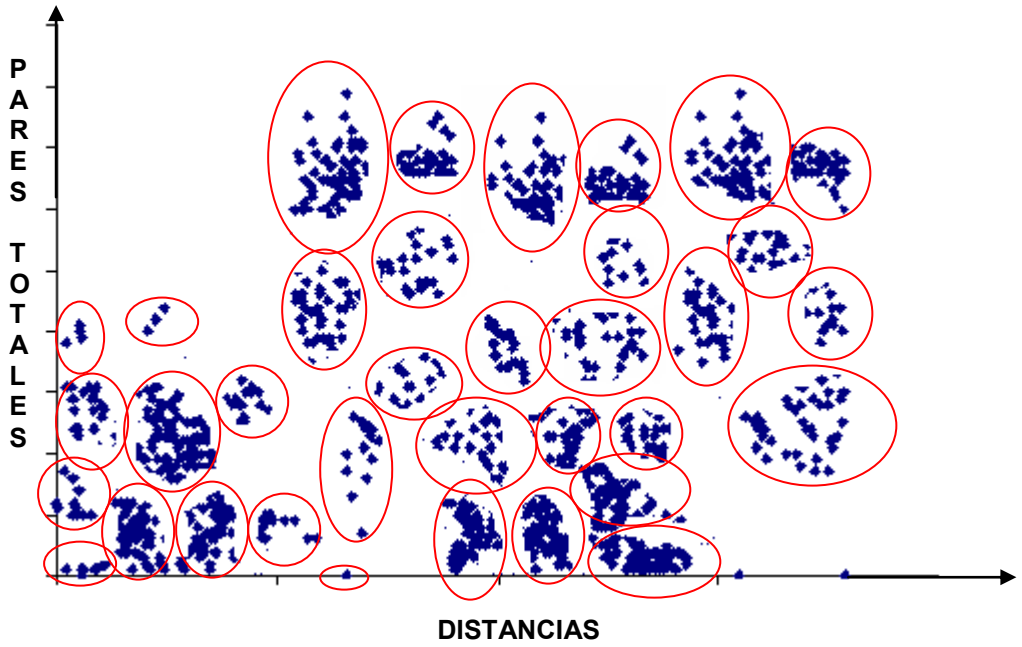
1.- A través del presente estudio se ha determinado en que consiste el Data Mining, las técnicas que existen y la utilización de las mismas.

2.- Las técnicas de la Minería de Datos o Data Mining son de mucha ayuda para poder extraer información representativa de una gran cantidad de registros.

3.- En un conjunto de n registros, donde n es grande, y se quiere aplicar las técnicas de minería de datos para calcular la distancia de edición, es recomendable realizarla con una comparación entre cadenas de caracteres, todos contra todos, debido a que de manera secuencial no se toman todos los registros, aunque es confiable, se pueden obtener mejores resultados con una comparación de todos contra todos.

4.- Si queremos decidir si un punto pertenece o no a un grupo necesitaremos basarnos en una medida de similitud o proximidad. Se han estudiado y sugerido un gran número de tales medidas, pero probablemente las más comúnmente usadas son las medidas de distancia de edición y en particular la distancia de Levenshtein.

Hacia esto:



Grupos categóricos que representen a la población total, es decir, a los 1857 registros.

5.- La aplicación de una de las técnicas de muestreo, como el muestreo aleatorio simple, nos permite proyectar nuestros resultados y conclusiones sobre nuestra población total, para disminuir el nivel de dificultad que tendríamos para hallar una distancia de edición óptima y definir nuestro criterio de discriminación.

Como conclusión final, podemos decir que la combinación de las técnicas de muestreo y datamining manejando distancias de edición, nos dan como resultado un mecanismo para categorizar en pocos grupos a una cantidad extensa de registros.

ANEXOS

Código fuente del proceso de agrupación de las descripciones.

```
Private Sub Command3_Click()
frmclasificacion.Show
Dim i As Integer
Dim ind_grupo As Integer
frmclasificacion.lstclasificacion.ListItems.Clear
frmclasificacion.lstclasificacion.ColumnHeaders.Clear

Dim new_column As ColumnHeader
Dim contador As Double
contador = conexion.conexion.Execute("select count(distancia) as maximo
from parcompleta where distancia>0")!maximo
Set new_column = frmclasificacion.lstclasificacion.ColumnHeaders.Add(, ,
"GRUPO 1")
new_column.Tag = 1
ind_grupo = 1

For i = 1 To contador
  frmclasificacion.lstclasificacion.ListItems.Add , , " "
```

```
Next i

'For i = 1 To lstdatos.ListItems.Count

Dim rsdatos As New ADODB.Recordset
Set rsdatos = conexion.conexion.Execute("select * from parcompleta")

Do While Not rsdatos.EOF
 If rsdatos!distancia <= min_distancia Then
  If ind_grupo = 1 Then

frmclasificacion.lstclasificacion.ListItems(VBA.Val(new_column.Tag)).Text
= rsdatos!distancia
    new_column.Tag = VBA.Str(VBA.Val(new_column.Tag) + 1)
  Else
frmclasificacion.lstclasificacion.ListItems(VBA.Val(new_column.Tag)).SubIt
ems(ind_grupo - 1) = rsdatos!distancia
    new_column.Tag = VBA.Str(VBA.Val(new_column.Tag) + 1)
  End If
 Else
  'agregar grupo
```

```
    Set new_column = frmclasificacion.lstclasificacion.ColumnHeaders.Add(,
, "GRUPO " & ind_grupo + 1)
    new_column.Tag = 1
    ind_grupo = ind_grupo + 1


frmclasificacion.lstclasificacion.ListItems(VBA.Val(new_column.Tag)).SubIt
ems(ind_grupo - 1) = rsdatos!distancia
    new_column.Tag = VBA.Str(VBA.Val(new_column.Tag) + 1)

  End If
rsdatos.MoveNext
Loop
'Next i
End Sub
```

Lista de artículos

A continuación les presentamos una breve lista de los artículos que fueron considerados para el proceso de agrupación.

DESCRIPCIONES DE PRODUCTOS
Abrecartas Aguila Plata de Ley
Abrecartas Brillo-mate
Abrecartas Escudo Chapado Plata
Abrecartas Mate
Absolut Azul 750 ml.
Absolut Citron 750 ml.
Absolut Mandrin 750 ml.
ACCESORIOS D TUBERIA (X EJ. EMPALMES, CODOS O MANGUITOS),D COBRE REFINADO
ACCESORIOS D TUBERIA MOLDEADOS D FUNDICION NO MALEABLE
ACCESORIOS DE TUBERIA DE PLASTICO (X EJ.JUNTAS, CODOS, EMPALMES (RACORES))
ACEITES CRUDOS DE PETROLEO O DE MINERAL BITUMINOSO
ACER ASPIRE 3002 LCI (S2.8/256
ACER INTEL TM-2355LCI CM 360
ACER INTEL TM-4652-LMI PM740
ACER TRAVELMATE 4100WLMI
ACER TRAVELMATE 4150 LCI
Agenda Avant
Agenda Naturbox
Agenda Sedalina
Agenda Semanal Grande Piel Vegetal
Agenda Semanal Grande Polipiel
Agenda Vaquetilla
BALASTOS (REACTANCIAS) P LAMPARAS O TUBOS D DESCARGA
BALLESTAS Y SUS HOJAS, D HIERRO O ACERO
BANDEJA 24*15 REP.
BANDEJA 33*23 POP.
BANDEJA 33*23 REP.
Bandeja Vacía-Bolsillos piel
BAÑADORES D PUNTO D FIBRAS SINTETICAS, P MUJERES O NIÑAS
BAÑADORES D PUNTO D LAS DMS MATERIAS TEXTILES, P MUJERES O NIÑAS
BARRAS D Fe O ACERO SN ALEAR CN MUESCAS,CORDONES,RELIEVES PRODUCIDOS EN LAMINADO
BARRAS Y PERFILES D COBRE REFINADO
Barton & Guestier Bordeaux Red - Francés 750 ml.
Barton & Guestier Bordeaux White - Francés
Barton & Guestier Cabernet Sauvignon - Francés 750 ml.
Barton & Guestier Chablis - Francés 750 ml.
Barton & Guestier Chardonnay-Francés 750 ml.
Barton & Guestier Cotes Du Rhone - Francés 750 ml.
Barton & Guestier Merlot - Francés 750 ml.
Barton & Guestier Rose D Anjou - Francés 750 ml.
Barton & Guestier Saint Emilion - Francés 750 ml.
BATTERY PACK CNX2-48V
BATTERY PACK CNX2-72V

Baúl Escritura Antigua
BAULES, MALETAS(VALIJAS)Y MALETINES,INCL.D ASEO,CN SUPERF EXTER CUERO NATURAL
BAULES,MALETAS(VALIJAS)Y MALETINES,INCL LS D ASEO,CN SUP EXTER D PLAST O MATER TEXTIL
BELKIN ADAPTADOR ENCHUFE P/VIAJE
BELKIN ADAPTADOR USB A ETHERNET 10/100
Bolígrafo 2 en 1
Bolígrafo Amarillo Bolita
Bolígrafo Amarillo Transp. Acid
Bolígrafo Amarillo-Niquel
Bolígrafo Azul Bolita
Bolígrafo Azul Carrera
Bolígrafo Azul Caucho
Bolígrafo Azul Escudo Transparente
Bolígrafo Azul Puntera Antideslizante
Bolígrafo Azul Transp. Acid
Bolígrafo Azul Transp. Rocket
Bolígrafo Azul-Niquel
Bolígrafo Capucha Negro
Bolígrafo Capucha Rojo
Bolígrafo de Madera
Bolígrafo Gord Azul
Bolígrafo Gord Rojo
Bolígrafo Gord Verde
Bolígrafo Granate Metalizado
Bolígrafo Hor
Boligrafo Madera Laser
Bolígrafo Madera-Niquel
Bolígrafo Morado Transp. Acid
Bolígrafo Naranja Transp. Acid
Bolígrafo Negro Escudo
Bolígrafo negro plata escudo
Bolígrafo Negro transp. Rocket
Bolígrafo Negro-Aluminio Ugr
Bolígrafo Plas Amarillo
Bolígrafo Plas Azul
Bolígrafo Plas Fusia
Bolígrafo Plas Verde
Bolígrafo Plas Verde Agua
Bolígrafo Plata Fino
Bolígrafo Plata Gord
Bolígrafo Plata Ugr
Caja Taracea 10 Cm.
Caja Taracea 12 Cm.
Caja Taracea 20 Cm.

Caja Taracea Portaboligrafos
Caja Taracea Portallavero
Caja Taracea Portamedalla Circular
Caja Taracea Portamedalla Rectangular
CAMARA DIG. SONY S60 4.1MP CYBERSHOT
CAMARA DIG.MICROTEK S1 3.0 MEGA LCD1.5"
CAMARA WEB ANIME KID SKYDIVER PINNACLE 640X480
CAMARA WEB COCKER SPANIEL 640X480
CAMARA WEB DOG SCHNAUZER 640X480
CAMARA WEB GENIUS GE-111 USB 640X480
CAMARA WEB GENIUS NB
CAMARA WEB HAMSTER 640X480
CAMARA WEB LOGITECH MESSENGER 640X480
CAMARA WEB LOGITECH QUICKCAM EXPRESS 640X480
CAMARA WEB PENGUIN 640X480
DISCO DURO 120 GB W.DIGITAL 7200 RPM
DISCO DURO 120 GB W.DIGITAL SATA 7200 RPM
DISCO DURO 160 GB SEAGATE 7200 RPM
DISCO DURO 160 GB W.DIGITAL 7200 RPM
DISCO DURO 200 GB W.DIGITAL 7200 RPM
DISCO DURO 40 GB W.DIGITAL 7200RPM
DISCO DURO 40 GB W.DIGITAL SATA RECERT. 7200 RPM
DISCO DURO 60 GB SEAGATE 5400 RPM REFURB.
DISCO DURO 80 GB HITACHI 7200 RPM
DISCO DURO 80 GB W.DIGITAL SATA RECERT.7200 RPM
DMS APARAT RESPIRATORIOS Y MASCARAS ANTIGAS,EXCEP MASCARAS D PROTECCION
DMS APARATOS D EMPALME O CONEXION P UNA TENSION < O = 260 V E INTENSIDAD < O = 30 A
DMS APARATOS D RAYOS X, P USO ODONTOLOGICO
DMS APARATOS E INSTRUMENT P PESAR CON CAPACIDAD > A 30 kG PERO <= A 5.000 kG
DMS APARATOS E INSTRUMENTOS P PESAR CON CAPACIDAD < O IGUAL A 30 kG.
DMS APARATOS ELECTRICOS D ALUMBRADO
DMS APARATOS MECANICOS P PROYECTAR,DISPERSAR MATERIAS LIQUIDAS O EN POLVO
DMS APARATOS P CORTE,SECCIONAMIENTO,PROTECCION,EMPALME, P TENSION <=1000V
DMS APARATOS P FILTRAR O DEPURAR LIQUIDOS
DMS APARATOS Y DISPOSITIV P TRATAMIENTO MATERIAS Q IMPLIQUEN CAMBIO TEMPERATURA
DMS ARTICULOS D CAMA Y ARTICULOS SIMILARES,BIEN CON MUELLES(RESORTES),O GUARNECIDOS
DMS ARTICULOS D GRIFERIA Y ORGANOS SIMILARES
DMS ARTICULOS D USO DOMESTICO Y SUS PARTES D FUNDICION, HIERRO O ACERO
DMS ARTICULOS D USO DOMESTICO Y SUS PARTES D FUNDICION, SIN

ESMALTAR
DMS ARTICULOS DE CONFITERIA SN CACAO (INCLUIDO EL CHOCOLATE BLANCO)
DMS ARTICULOS DE USO DOMESTICO Y ARTIC D HIGIENE O D TOCADOR, DE PLASTICO
DMS ARTICULOS P FIESTA,CARNAVAL U OTRAS DIVERSIONS,INC LS D MAGIA Y ARTICULOS SORPRESA
DMS ARTICULOS P JUEGOS D SOCIEDAD
DMS ARTICULOS PARA TRANSPORTE O ENVASADO, DE PLASTICO
DMS ARTICULOS ROSCADOS:DE FUNDICION, HIERRO O ACERO
EPSON STYLUS C45UX 2CART ADIC
EPSON STYLUS C65
EPSON STYLUS CX1500 CART.BK/RE
EPSON STYLUS CX3500 CART.BK/RE
EPSON STYLUS CX4500 CART.BK/RE
EPSON STYLUS PHOTO R200M
EPSON STYLUS PHOTO R300M
Ernest & Julio Gallo Reserva
Estuche de cuño
Estuche de Madera 1 Pza.
Estuche de Madera 2 Pzas.
Estuche de Pinceles
Estuche de Plumilla
Estuche Manicura Aluminio
Estuche Rectangular 6 lapices
Estuche Rosado 11*22 cm
GENIUS HEADSET HS-02F-FOLDABLE
GENIUS HEADSET HS-02N-NEK TYPE
GENIUS HEADSET HS-03A (EAR)
GENIUS JOYSTICK F-16
GENIUS JOYSTICK F23 U/NEW
GENIUS JOYSTICK F-31U VIBRAT.
GENIUS JOYSTICK F2000 F23 DIG.
GENIUS MOUSE NET VALUE PS2
GENIUS MOUSE EASYPRO PS2 S5
GENIUS MOUSE EASYPRO SERIAL S5
GENIUS MOUSE EYE V2 PS2 BEIGE
GENIUS MOUSE EYE V2 PS2 METALI
GENIUS MOUSE MINITRAVELER U+P
GENIUS MOUSE NET ROLLER PS2
GENIUS MOUSE NETSC PS2 BLACK
GENIUS MOUSE NS+SUP.U+P SILVE
GENIUS MOUSE NS+TRAV.PS2 BLACK
GENIUS MOUSE NS+TRAV.PS2 BLUE
GENIUS MOUSE NS+TRAV.PS2 METAL
GENIUS MOUSE NS+TRAV.U+P BLACK
GENIUS MOUSE NS+TRAV.U+P BLUE

GENIUS MOUSE NS+TRAV.U+P METAL
GENIUS MOUSE NS+TRAV.U+P RUBY
GENIUS MOUSE OPT V2 NETEYE PS2
GENIUS MOUSE OPTICO NETEYE PS2
GENIUS MOUSE WIRELESS POWERS
GENIUS MOUSE XSCROLL PS2 BLACK
GENIUS VIDEO CAM MESSENGER
GENIUS VIDEO CAM TREK BLACK
GENIUS VIDEO CAM WEB
GENIUS VIDEOCAM EXPRESS V2 100
GENIUS VIDEOCAM NB K971C007F
GENIUS VIDEOCAM SLIM BLUE
HD IDE 20GB 5400 RPM MAXTOR
HD IDE 20GB 7200 RPM MAXTOR
HD IDE 30GB 7200 RPM MAXTOR
HD IDE 40GB 7200 RPM SANSUNG
HD IDE 40GB 7200 RPM WDIGITAL
HD IDE 80GB 7.2 RPM WD S-ATA
HD IDE 80GB 7200 RPM HITACHI
HD IDE 80GB 7200 RPM SANSUNG
HD IDE 80GB 7200 RPM WDIGITAL
HD IDE 80GB 7200/8 MB WDIGITAL
K-BYTE ESPUMA LIMPIADORA 120ML
K-BYTE ESPUMA LIMPIADORA 400ML
K-BYTE LIMPIA PANTALLAS 120ML
K-BYTE LIMPIA PANTALLAS 400ML
K-BYTE LIMPIEZA COMPACT DISC
K-BYTE REPARADOR DISCOS COMPACTOS (CD)
K-BYTE SET LIMPIA PANTALLA LCD,TFT CRISTAL LIQUIDO
KIT ATX336-02M.TWR NET+MULT+SP
MB BIOSTAR S478 P4TSED2 I865PE
MB BIOSTAR S478 P4VMA-M PRESCO
MB BIOSTAR S478 U8668DV7P
MB BIOSTAR S754 K8VGAM
MB BIOSTAR S775 P4M80-M7
MB BIOSTAR S-A KM400 M7VIZ 5.1
MB BIOSTAR S-A M7VIG400
MB BIOSTAR S-A M7VIG400+2200
MB BIOSTAR S-A M7VIG400+3200
MB INTEL S478 D845GVSR BOX
MB INTEL S478 D845GVSRL BOX
MB INTEL S478 D865GBF BOX
MB INTEL S478 D865GVHZ BOX
MB INTEL S478 D865PERL BOX
MB INTEL S478 SERVER S875WP1E
MB INTEL S775 D915 GAGL BOX

MB INTEL S775 D915 GLVGL BOX
MB INTEL S775 D915 GVWBL BOX
MEM DDR 128 MB 333 MHZ KINST.
MEM DDR 128 MB 333 MHZ MARCA
MEM DDR 1GB MB 400 MHZ MARCA
MEM DDR 256 MB 333 SPK NBOOK
MEM DDR 256 MB 333 MHZ KINST.
MEM DDR 256 MB 333 MHZ MARCA
MEM DDR 256 MB 400 SPECTEK
MEM DDR 256 MB 400 MHZ KINST.
MEM DDR 256 MB 400 MHZ MARCA
MEM DDR 512 MB 333 MARC NBOOK
MEM DDR 512 MB 333 MHZ KINST.
MEM DDR 512 MB 333 MHZ MARCA
MEM DDR 512 MB 400 MHZ KINST.
MEM DDR 512 MB 400 MHZ MARCA
MEM DDR2 256 MB 333 MHZ MARCA
Monedero Billetero Cremallera
Monedero Billetero Foam Granate
Monedero Billetero Foam Kaki
Monedero Billetero Mod.49
Monedero Billetero Mod.52
Monedero Billetero Sedalina
Monedero Billetero Sedalina Habana
Monedero Euro
Monedero Foam Granate
Monedero Foam Kaki
Monedero Naturbox
Monedero Naturbox negro
Monedero Vaquetilla
MONITOR SAMSUNG 15" 591S BEIGE
MONITOR SAMSUNG 15" 591S NEGRO
MONITOR SAMSUNG 17" 793DF FLAT BEIGE
MONITOR SAMSUNG 17" 793V BEIGE
MONITOR SAMSUNG 17" 793V NEGRO
MONITOR SAMSUNG 17" TFT 710N
MONITOR SAMSUNG 19" 997MB
MOUSE GENIUS MINI OPT. TRAVELER SILVER
MOUSE GENIUS MINI OPT. TRAVELER WHITE
MOUSE GENIUS NETSCROLL PS/2 BLACK
MOUSE GENIUS NETSCROLL USB
MOUSE GENIUS OPT. PS/2 NETSCROLL BEIGE
MOUSE GENIUS OPT. TRAVELER BLACK
MOUSE GENIUS OPT. TRAVELER BLUE
MOUSE GENIUS OPT. TRAVELER RUBY
MOUSE GENIUS OPT. TRAVELER SILVER

MOUSE GENIUS USB WIRELESS POINTER+LASER
MOUSE GENIUS VALUE PS/2
MOUSE LOGITECH OPTICO FOOTBALL USB
MOUSE LOGITECH OPTICO MX-310 PS/2 USB
MOUSE LOGITECH OPTICO MX-510 BLUE PS/2 USB
MOUSE LOGITECH OPTICO MX-510 RED PS/2 USB
MOUSE LOGITECH OPTICO PS/2 USB/GRIS/IVORY
PLATO 12 CM. POP.
PLATO 12 CM. REP.
PLATO 20 CM. POP.
PLATO 20 CM. REP.
PLATO 23 CM. POP.
PLATO 23 CM. REP.
PLATO 30 CM. POP.
PLATO 30 CM. REP.
TECLADO GENIUS KB06 BLACK PS/2
TECLADO GENIUS MINI MULTIMEDIA USB
TECLADO GENIUS MULTIMEDIA KB-29E
TECLADO GENIUS MULTIMEDIA PS/2 BEIGE
TECLADO GENIUS MULTIMEDIA USB KB-12E
TECLADO GENIUS PS/2 BEIGE
TECLADO LOGITECH DELUXE PS/2
TECLADO LOGITECH INTERNET BLANCO PS/2
TECLADO LOGITECH+MOUSE PS/2
TECLADO MAXXTRO USB BEIGE
TECLADO MULTIMEDIA PS/2 BEIGE
TECLADO MULTIMEDIA PS/2 KB-1616 BEIGE
TECLADO MULTIMEDIA PS/2 KB-1616 NEGRO
TECLADO MULTIMEDIA PS/2 NEGRO
TECLADO MULTIMEDIA USB BLUE/IVORY-2616
TECLADO NUMERICO USB P/NOTEBOK
TECLADO PCTRONIX MULTIMEDIA KB-3618 NEGRO
TECLADO PS/2 KB-1607 BEIGE
TECLADO PS/2 KB-1607 NEGRO
TECLADO PS/2 KB-1906 BLACK/SILVER
TECLADO PS/2 NEGRO
TECLADO SUNSHINE MULTIM. PS2
TEJ BLANQUEADS ALG >=85% LIGAMENTO SARGA,INC CRUZADO CURSO <= 4, GRAMAJE >200G/M2
TEJ ID ESTAMP CN UN CONTEN ALG >=85% LIGAMENTO TAFETAN GRAMAJE >100G/M2, <=200 G/M2
TEJAS D CERAMICA
TEJID CN HILADOS DIST COLOR FIB POLIEST<85% LIGAM TAFETAN MEZC EXCL ALGOD GRAM<=170G/M2
TEJID CN HILADOS DISTINT COLORES,CONT ALG >085%,D LIGAMENTO TAFETAN,GRAMAJE>100G/M2
TEJID ESTAMPADO ALG>=85% LIGAMENTO SARGA,INC CRUZADO CURSO <=

A4,GRAMAJE <=200G/M2
TEJID ESTAMPADOS FB DISCONT POLIEST <85% LIGA TAFETA MEZC EXCL ALGOD GRAM<=170G/M2
TEJIDO D MEZCLILLA("DENIM")CN HILADOS D DISTINT COLORES,ALG>=85%,GRAMAJE >200G/M2
TEJIDO TEÑIDO,CONTEN ALGODON >=85% LIGAMENTO SARGA,INC CRUZADO CURSO <=4,GR<=200G/M2
TELAS IMPREGNADAS,RECUBIERTAS,REVESTIDAS O ESTRATIFICADAS CN POLICLORURO D VINILO
Termómetro Alcoholímetro (Para Vinos)
Termómetro madera Alcoholímetro (Para Vinos)
Terrazas de los Andes
TIJERAS Y SUS HOJAS
TONER REFILL HP 2100 C4096A
TONER REFILL HP3P 92275A
TONER REFILL HP4/5000 C4127X
TONER REFILL HP4V C3900A
TONER REFILL HP5M 92298A
TONER REFILL HP5SI C3909A
TONER REFILL HP6L/1100 C3906A
TONER REFILL HP6P C3903F
TONER XEROX 3116
TOSHIBA SAT. P35-SP611
TOSHIBA SAT.A70-SP211
TOSHIBA SAT.L10-SP104
TOSHIBA SAT.M30X-SP111
TOSHIBA TECRA A3-SP611
TRANFORMADORES D DIELECTRICO LIQUIDO, D POTENCIA > A 10.000 kVA
TRANSFORMADORES D DIELECTRICO LIQUIDO, D POTENCIA >A 10kVA PERO <O = 650kVA
T-SHIRTS Y CAMISETAS INTERIORES, D PUNTO: DE ALGODON
TUBO PLEYADE 40 CM. CON 150 GR. FLOR
TUBO PLEYADE 50 CM. CON 300 GR. FLOR
TUBOS RIGIDOS DE POLIMEROS DE CLORURO DE VINILO
UPS 1200 VA C/REGULADOR TRV
UPS 1500 VA C/REGULADOR TRV
UPS 2000 VA C/REGULADOR TRV
UPS 625 VA C/REGULADOR TRV
UPS 900 VA C/REGULADOR TRV
UPS LIEBERT GXT2U 1000
UPS LIEBERT GXT2U 1500
UPS LIEBERT GXT2U 2000
UPS LIEBERT GXT2U 3000
UPS LIEBERT PA 500
UPS LIEBERT PA 650
UPS LIEBERT PS 3000
UPS LIEBERT PS2 1440

UPS LIEBERT PS2 2200
UPS LIEBERT PSP XT 1250
UPS LIEBERT PSP XT 450
UPS LIEBERT PSP XT 700
UPS SEEBEK PS1 500VA 25 MIN
USB PEN DRIVE 128MB.2.0
USB PENDRIVE 1GB MB 2.0
USB PENDRIVE 256 MB 2.0
USB PENDRIVE 2GB MB 2.0
USB PENDRIVE 512 MB 2.0
VAJILLA Y DMS ARTICULOS PARA EL SERVICIO DE MESA O DE COCINA, DE PLASTICO
VALVULAS D ALIVIO O SEGURIDAD.
VALVULAS D COMPUERTA D DIAMETRO NOMINAL <=100MM, P PRESIONES >=A 13.8 Mpa
VASO WHISKY POP.
VASO WHISKY REP.
VENTILADOR CPU AMD ATHLON
VENTILADOR PARA SLOT CASE
VENTILADORES D MESA,PIE,PARED,TECHO,O VENTANA CN MOTOR ELECTR INCORP POTENC <=125W
VESTIDOS D FIBRAS SINTETICAS, P MUJERES O NIÑAS
VGA ATI RADEON 7000 32DDR
VGA ATI RADEON 7000 64DDRTV
VGA ATI RADEON 9200SE 128DDRTV
VGA ATI RADEON 9200SE 256DDRTV
VGA ATI RADEON 9200SE 64DDRTV
VGA ATI RADEON 9250 256DDRTV
VGA ATI RADEON 9250SE 128DDRTV
VGA ATI RADEON 9550SE 128DDRTV
VGA ATI RADEON 9550SE 256DDRTV
VGA ATI RADEON 9600 256DDRTV
VGA ATI RADEON 9600SE 128DDRTV
VGA BIO. MX4000 128 DDR TV
VGA BIO. MX4000 64 DDR TV
VGA BIO.FX-5200 128 DDR TV 8X
VGA BIO.FX-5700LE 128 DDR TV8X
VGA BIO.FX-5700LE 256 DDR TV8X
VGA GEF-4 128MB MX4000TV DDR8X
VGA GEF-4 64MB MX4000TV DDR8X
VGA GEF-4 BIOST 64MB440TVDDR8X
VGA MSI 5700LE TD128
VGA MSI 5700LE TD256
VGA MSI ATI 9550 128M DDR TV D
VGA MSI ATI R9200 128M DDR TV
VGA MSI ATI RX9200SE T128 DDR
VGA MSI ATI RX9250 128M DDR TV

VGA MSI FX5200T128 128DDR TV
VGA MSI FX5200TD128LE 128DR TV
VGA MSI FX5200TD128LF 128DR TV
VGA MSI FX5500T128 128DDR TV
VGA MSI FX5500TD256 256DDR TV
VGA MSI MX4000 T128 128DDR TV
VGA MSI MX4000 T64 DDR TV
VGA MSI NX6200 AGP TD128MB TV
VGA MSI NX6200 PCI-E TD128MB
VGA MSI NX6200AX 64B TD128MBTV
VGA MSI NX6600 AGP VTD256MB TV
VGA MSI RX600XT- TD128E PCI-E
VGA MSI RX9550 AGP TD128MB TV
VGA MSI RX9600 PRO TD128MB TV
VIDEO ALHAMBRA
VISILLOS Y CORTINAS;GUARDAMALLETAS Y RODAPIES D CAMA,D PUNTO D FIBRAS SINTETICAS
White Zifandel - Californiano 750 ml.
WIRELESS LG 11G ACCES POINT
WIRELESS LG 11G ADAPT.PCMCIA
WIRELESS MSI 11G ACCES POINT
WIRELESS MSI 11G ADAPTADOR PCI
WIRELESS MSI 11G ADAPTADOR USB

Grupos categorizados

ROLLOS
ROLLOS PARA FAX NORMA
ROLLOS PARA SUMADORA # 37 NORMA
ROLLOS PARA SUMADORA # 57 NORMA
ROLLOS PARA SUMADORA # 70 NORMA

PAPEL CARBON
PAPEL BOND BASE 20 P/FOTOCOP. OFICIO - ALPES
PAPEL CARBON KORES RESMA CARTA
PAPEL CARBON KORES RESMA OFICIO
PAPEL CARBON NORMA RESMA CARTA

TECLADO GENIUS MULTIMEDIA
TECLADO GENIUS KB06 BLACK PS/2
TECLADO GENIUS MINI MULTIMEDIA USB
TECLADO GENIUS MULTIMEDIA KB-29E
TECLADO GENIUS MULTIMEDIA PS/2 BEIGE
TECLADO GENIUS MULTIMEDIA USB KB-12E

TECLADO MULTIMEDIA PS/2 BEIGE
TECLADO LOGITECH+MOUSE PS/2
TECLADO MAXXTRO USB BEIGE
TECLADO MULTIMEDIA PS/2 BEIGE
TECLADO MULTIMEDIA PS/2 KB-1616 BEIGE
TECLADO MULTIMEDIA PS/2 KB-1616 NEGRO

TONER REFILL HP
TIJERAS Y SUS HOJAS
TONER REFILL HP 2100 C4096A
TONER REFILL HP3P 92275A
TONER REFILL HP4/5000 C4127X
TONER REFILL HP4V C3900A
TONER REFILL HP5M 92298A
TONER REFILL HP5SI C3909A
TONER REFILL HP6L/1100 C3906A
TONER REFILL HP6P C3903F

TOSHIBA SAT.
TONER XEROX 3116
TOSHIBA SAT. P35-SP611
TOSHIBA SAT.A70-SP211
TOSHIBA SAT.L10-SP104
TOSHIBA SAT.M30X-SP111

UPS 2000 VA C/REGULADOR TRV
TUBOS RIGIDOS DE POLIMEROS DE CLORURO DE VINILO
UPS 1200 VA C/REGULADOR TRV
UPS 1500 VA C/REGULADOR TRV
UPS 2000 VA C/REGULADOR TRV
UPS 625 VA C/REGULADOR TRV
UPS 900 VA C/REGULADOR TRV

UPS LIEBERT PA 650
UPS LIEBERT GXT2U 1000
UPS LIEBERT GXT2U 1500
UPS LIEBERT GXT2U 2000
UPS LIEBERT GXT2U 3000
UPS LIEBERT PA 500
UPS LIEBERT PA 650
UPS LIEBERT PS 3000
UPS LIEBERT PS2 1440
UPS LIEBERT PS2 2200
UPS LIEBERT PSP XT 1250
UPS LIEBERT PSP XT 450
UPS LIEBERT PSP XT 700

USB PENDRIVE 256 MB 2.0
UPS SEEBEK PS1 500VA 25 MIN
USB PEN DRIVE 128MB.2.0
USB PENDRIVE 1GB MB 2.0
USB PENDRIVE 256 MB 2.0
USB PENDRIVE 2GB MB 2.0
USB PENDRIVE 512 MB 2.0

VGA ATI RADEON 9250SE 128DDRTV
VESTIDOS D FIBRAS SINTETICAS, P MUJERES O NIÑAS
VGA ATI RADEON 7000 32DDR
VGA ATI RADEON 7000 64DDRTV
VGA ATI RADEON 9200SE 128DDRTV
VGA ATI RADEON 9200SE 256DDRTV
VGA ATI RADEON 9200SE 64DDRTV
VGA ATI RADEON 9250 256DDRTV
VGA ATI RADEON 9250SE 128DDRTV
VGA ATI RADEON 9550SE 128DDRTV
VGA ATI RADEON 9550SE 256DDRTV
VGA ATI RADEON 9600 256DDRTV
VGA ATI RADEON 9600SE 128DDRTV
VGA BIO. MX4000 128 DDR TV
VGA BIO. MX4000 64 DDR TV
VGA BIO.FX-5200 128 DDR TV 8X

VGA GEF-4 64MB MX4000TV DDR8X
VGA BIO.FX-5700LE 256 DDR TV8X
VGA GEF-4 128MB MX4000TV DDR8X
VGA GEF-4 64MB MX4000TV DDR8X
VGA GEF-4 BIOST 64MB440TVDDR8X

VGA MSI FX5500T128 128DDR TV
VGA MSI 5700LE TD256
VGA MSI ATI 9550 128M DDR TV D
VGA MSI ATI R9200 128M DDR TV
VGA MSI ATI RX9200SE T128 DDR
VGA MSI ATI RX9250 128M DDR TV
VGA MSI FX5200T128 128DDR TV
VGA MSI FX5200TD128LE 128DR TV
VGA MSI FX5200TD128LF 128DR TV
VGA MSI FX5500T128 128DDR TV
VGA MSI FX5500TD256 256DDR TV
VGA MSI MX4000 T128 128DDR TV
VGA MSI MX4000 T64 DDR TV
VGA MSI NX6200 AGP TD128MB TV
VGA MSI NX6200 PCI-E TD128MB

Cabe recalcar que esto es una muestra de todos los grupos que se formaron.

BIBLIOGRAFIA

1. (Michael A. J. Berry & Gordon S. Linoff - 2004), 2daEd. - Data.Mining Techniques for Marketing Sales and Customer Support.
2. (Ian H. Witten & Eibe Frank – 2003) - Data Mining Practical Machine Learning Tools and Techniques.
3. (César Pérez - 2000) - Técnicas de Muestreo Estadístico – Teoría, práctica y aplicaciones informáticas.
4. (John Freund – Maryless Miller – Irwin Miller - 2000) - Estadística Matemática con Aplicaciones, 6ta Edición.
5. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm (al 16-Jun-06).
6. http://www.monografias.com/trabajos26/data-mining/data-mining.shtml (al 19-Jul-06).
7. http://www.monografias.com/trabajos/datamining/datamining.shtml (al 10-Oct-06).
8. http://es.wikipedia.org/wiki/Distancia_de_Levenshtein (al 15-Sep-06).
9. www.ecci.ucr.ac.cr (al 15-Oct-06).