Análisis de Grandes Volúmenes de Datos
Carlos Espinoza Peraza - B92786
Reporte 15 de julio: Clustering

Resumen de los temas estudiados

Flujos de Datos

Usualmente se realizan estudios sobre datos estáticos, que se encuentran almacenados, no tienen límite de tiempo para ser procesados, se aplican técnicas para mayor entendimiento y no generan un efecto inmediato. El análisis de un flujo de datos, a diferencia de estos tipos de análisis, permite el procesamiento en un menor tiempo, sin necesidad de almacenamiento

Stream Processing

- Permite el procesamiento sin almacenar
- Procesamiento en tiempo real
- No requiere de un tiempo de espera para que un algoritmo entrene un modelo
- Diseñado para trabajar con `datasets infinitos`
- Recolecta, integra y analiza ráfagas de datos

Cardinalidad:

> Bounded: tamaño finito, inicio y fin, posible almacenarla, proc Batch

> Unbounded: tamaño infinito, registros pequeños de alto volúmen, dist inconsistente, no respeta secuencia

Constitución:

> Tablas: SQL, NoSQL, Archivos, Datos en reposo

> Streams: datos fluyen en el tiempo, datos en movimiento

> **Conceptos**

Característica	Batch	Streaming
Frecuencia	Poco frecuente	Ejecución permanente
Desempeño	Alta latencia	Baja latencia
Fuentes de datos	DBs, APIs, Files	Colas de mensajes, eventos, transacciones
Tipo de análisis	Complejo	Simple

Característica	Batch	Streaming
Procesamient	Después de almacenar	Primero se procesa

> **Sistemas en tiempo real**

Clasificación	Latencia	Tolerancia a retraso
Hard	Microsegundos - milisegundos	Ninguna
Soft	Milisegundos - segundos	Baja tolerancia
Near	Segundos - minutos	Alta

> **Técnicas Utilizadas**

Técnica	Concepto
Time agnostic	Tiempo irrelevante, refleja un elemento filtrado
Approximation	Toman un dato y lo transforman a algo similar
Windowing	Partir un stream de datos continuos
W by Proc Time	Apilar datos durante tiempo `t`
W by Event Time	Reflejan el tiempo del evento
Tumbling Window	Se dispara cuando se llena una ventana
Filtros	Se quiere aceptar tuplas u observaciones del stream que cumplan una propiedad y descartar tuplas que no cumplan esta propiedad

Análisis de Grafos

- Modelo de grafos es una alternativa al modelo tradicional de DW
- Permiten integran datos de gran escala que provienen de distintas fuentes
- Permite el manejo de datos estructurados y no estructurados
- Permite las consultas de manera indirecta

Componentes de un grafo

> Vértices

> Aristas

Grafos como redes

- Permite representar formas de interacción entre entidades
- Permite detectar patrones

Algoritmos Para Análisis de Grafos

Tipo de Algoritmo	Concepto
Redes	Conectadas de manera particular y cercana
Caminos	Estudio de las formas y rutas que conectan las entidades
Segmentación	Examinan las propiedades de los vértices y aristas para identificar características útiles para agruparlos
Detección de patrones	Apilar datos durante tiempo t
W by Event Time	Encuentran anomalías que requieren investigación
Métricas de grafos	Asociadas con la red misma, grado de los vértices, centralidad y distancia entre vértices

Retos para el análisis

- Acceso a memoria impredecible
- Modelos de crecimiento
- Interacciones dinámicas
- Particionamiento complejo
- Medidas de similitud

Segmentación de Grafos

- Betweenness
- Top-down / Bottom-up
- Búsqueda Directa
- Grafo Bipartita Completo
- Búsqueda A Priori

Comentarios sobre la materia estudiada

El análisis de datos por medio de ráfagas tiene aplicaciones que juegan papeles importantes en la vida cotidiana, especialmente con el auge de técnologías 5G que permiten la conexión rápida con objetos que a su vez deben realizar operaciones rápidas de procesamiento para poder generar un efecto como consecuencia de la entrada recibida. El análisis de grafos permiten un análisis visual y ordenado de las diferentes interacciones que tienen distintas entidades

Dudas sobre la materia

> Ninguna

Posible uso como profesional

En la actualidad está de moda el uso de bases de datos NoSQL que implementan formas de almacenar y visualizar datos en forma de grafos. Un ejemplo de esto es Neo4j que permite el análisis de los datos almacenados en forma de grafo (relaciones).

Problemas que podría resolver con las técnicas estudiadas

Para el análisis de grafos se pueden tener los típicos problemas de optimización que permiten encontrar rutas más cortas entre elementos. Teniendo esto en cuenta, se podría combinar con un análisis de ráfagas para poder actualizar en tiempo real los pesos de arístas, lo cual permitiría poder tener resultados reales de lo que podría ser la mejor ruta para llegar de A a B tomando en cuenta el estado actual del tráfico