

DATOS MASIVOS I

UNIDAD 1 INTRODUCCIÓN

CONCEPTOS BÁSICOS

1 de Febrero de 2023

Para hoy



Conceptos básicos.



Principio de bonferroni

Almacenamiento



Consideraciones Generales



Centros de datos están en hiperexpansión. Utilizar la energía de bajo costo y los espacios inmobiliarios económicos, son dos objetivos que se perdiguen.



El almacenamiento de datos se centra en la persistencia, confiabilidad y durabilidad de los datos. Los objetivos deben ser facilidad, velocidad y rentabilidad.



Existe el problema de extraer valor de los datos que conducen a nuevos descubrimientos y abren nuevas oportunidades de investigación y comerciales.
¿Hay solución?



Bases de datos paralelas.



En una base de datos paralela es necesario distribuir los datos entre los diferentes nodos de la red.



El particionamiento de datos permite el procesamiento concurrente de transacciones y la paralelización de consultas, esto es importante por cuestiones de manejo, mantenimiento y rendimiento.



Además, alivia la carga de E/S y libera ancho de banda.



Bases de datos distribuidas.



Los datos se encuentran dispersos en diferentes sitios de la red, los cuales se comunican entre sí.



La distribución de los datos tiene ventajas como: disponibilidad, uso compartido de datos, fiabilidad y descentralización. ¿Hay diferencia con BD paralelas?



Hadoop.



HadoopDB usa Postgres como capa de base de datos en cada nodo.



Hadoop/MapReduce como capa de comunicación para coordinar todos los nodos, y



Hive como capa de traducción.

Hadoop



Hadoop.



Como resultado se tiene una base de datos paralela, en la cual es posible interactuar usando un lenguaje de tipo SQL.



El componente principal de HadoopDB es el framework Hadoop. Hadoop se compone de dos capas: almacenamiento (Sistema de Archivos Distribuido HDFS) y procesamiento de datos (framework MapReduce).



Principio de Bonferroni

¿Cuál o cuáles son los riesgos que se presentan en el análisis/minería de datos?



Un gran riesgo de la minería de datos es que se “*descubren*” patrones que no tienen sentido.



Meaningfulness of Answers

Si se buscan patrones
interesantes en más lugares
de los que admite su
cantidad de datos,

muy probablemente
se encontrarán
resultados sin
significancia

Si se buscan patrones interesantes en más lugares de los que admite su cantidad de datos,

muy probablemente se encontrarán resultados sin significancia. ¿Por qué?



Se calcula el número
esperado de ocurrencias de
un evento, bajo la suposición
que es aleatorio.



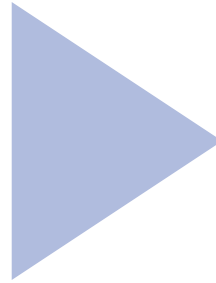
Si el número calculado es
mucho mayor al de las
ocurrencias reales,



entonces las conclusiones
que puedas sacar a partir de
estos eventos,
probablemente sean falsas.

Ejercicio

Supongamos que
creemos que ciertos
grupos de terroristas
se reúnen
ocasionalmente en
hoteles para tramar
hacer el mal.



Queremos encontrar
personas (no
relacionadas) que al
menos dos veces se
hayan hospedado en
el mismo hotel el
mismo día.

Datos del ejercicio

- Hay 1000 millones de personas
- Cada persona va a un hotel una vez cada 100 días
- Un hotel hospeda 100 personas y hay 100,000 hoteles (capaces de hospedar al 1 % del total de personas)
- Para detectar un terrorista buscamos pares de personas que en 2 días distintos en una ventana de 1000 días fueron al mismo hotel

Solución

- La probabilidad de que 2 personas decidan ir a un hotel cualquiera de los 100 días es $0.01 \times 0.01 = 0.0001$

- La probabilidad de que además elijan el mismo hotel es

$$\frac{0.0001}{10^5} = 10^{-9}$$

- La probabilidad de que 2 personas visiten el mismo hotel en 2 días distintos es $10^{-9} \times 10^{-9} = 10^{-18}$

Resultado

- El número total de posibles pares de personas es

$$\binom{10^9}{2} \approx 5 \times 10^{17}$$

- El número de pares de días es

$$\binom{1000}{2} \approx 5 \times 10^5$$

- Por lo tanto, el número esperado de personas que visitan el mismo hotel en 2 días distintos es

$$(5 \times 10^{17}) \times (5 \times 10^5) \times 10^{-18} = 250,000$$

Ejercicio

Detectar un terrorista buscando pares de personas que,

en dos días distintos,

en una ventana de 3000 días,

fueron al mismo hotel.

Solución

5×10^{17} → Número de pares de personas
 10^{-18} → Probabilidad de que dos personas se encuentren en un mismo hotel en dos días diferentes
 4.5×10^6 → Posibles pares de días en una ventana de 3000

$$= (5 \times 10^{17}) \times (4.5 \times 10^6) \times (10^{-18})$$

$$= 2,250,000$$

¿Cuáles son las conclusiones?

Finalmente

Digamos que hay 10 pares de terroristas que definitivamente se hospedaron en el mismo hotel dos veces.

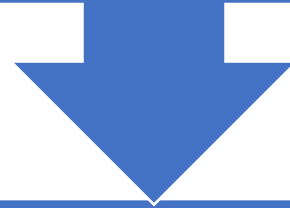


Los analistas tienen que examinar 250,000 candidatos para encontrar los 10 casos reales.

Respuesta: No va a pasar.

Finalmente

Supongamos que hay (digamos) 10 pares de terroristas que definitivamente se hospedaron en el mismo hotel dos veces.



Los analistas tienen que examinar 250,000 candidatos para encontrar los 10 casos reales.

Respuesta: No va a pasar.

¿Cómo podemos mejorar el esquema?

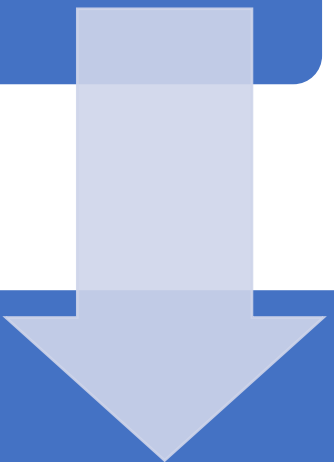
Cuando se busque una característica en particular,

por ejemplo, “dos personas se hospedaron dos veces en el mismo hotel”,

es mejor asegurarse de que la característica no permita tantas posibilidades,

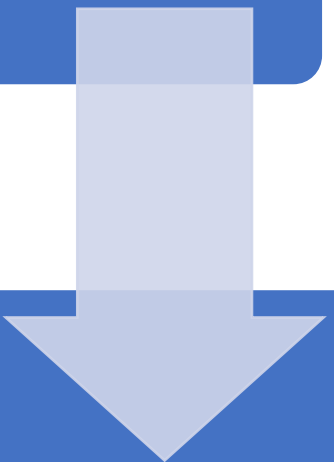
de manera que los datos aleatorios seguramente producirán hechos “de interés”.

Bonferroni discounting



Cuando se ejecutan k pruebas, un enfoque sencillo pero conservador es reducir el umbral de significación para cada prueba a $0.05/k$.

Bonferroni discounting



Esto garantiza que la probabilidad de que cualquiera de los resultados se produzca por casualidad será inferior a 0.05.

Rhine Paradox





Coffe
dementia



CEDARS-SINAI
SPINE CENTER

1-800-CEDARS-1
(1-800-233-2771)

Search Health 3,000+ Topics

For First Time, AIDS Vaccine Shows Some Success

By DONALD G. MCNEIL JR.
Published: September 24, 2009

Scientists said Thursday that a new [AIDS](#) vaccine, the first ever declared to protect a significant minority of humans against the disease, would be studied to answer two fundamental questions: why it worked in some people but not in others, and why those infected despite vaccination got no benefit at all.



The vaccine — known as RV 144, a combination of two genetically engineered vaccines, neither of which

☒ SIGN IN TO
RECOMMEND

TWITTER


 COMMENTS
(33)

 SIGN IN TO
E-MAIL

 PRINT

 REPRINTS

 SHARE

 SHARE

Well

Tara Parker-Pope on Health

**Tips for Navigating Medicare**

October 16, 2009

Show Off Your Vegetables With Pasta

October 16, 2009

High-Deductible Health Plans: Better for You or Your Employer?

October 16, 2009

The Roving Runner: Prospect Park

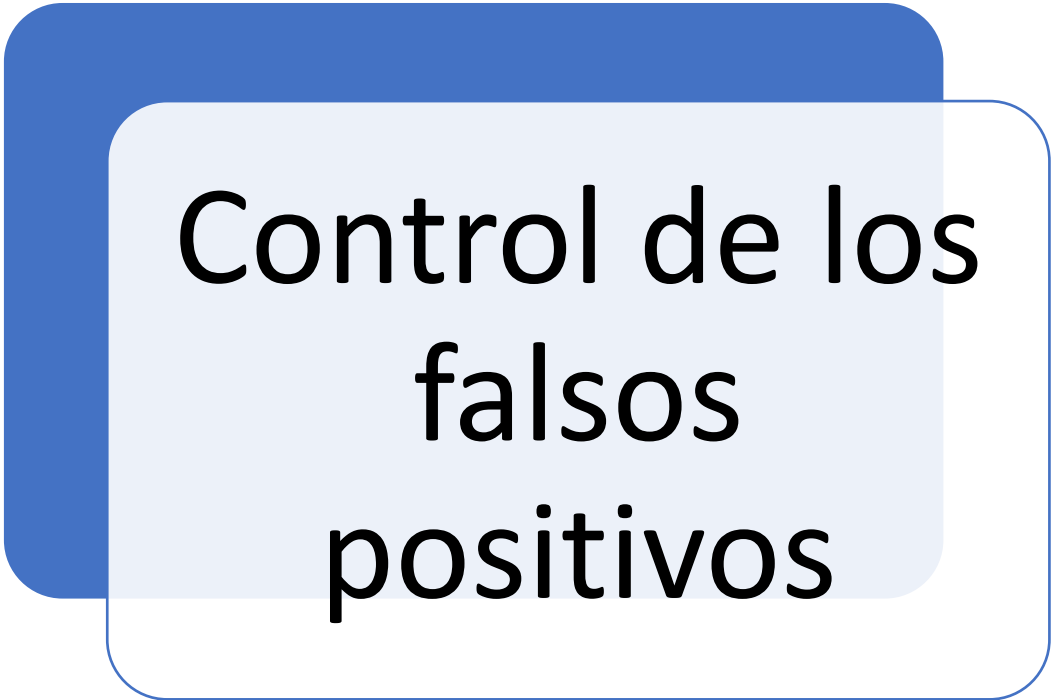

October 16, 2009

Alternative Medicine Cabinet: Thyme for Toenails

October 15, 2009

TicketWatch - Theater Offers by E-Mail

Sign up for ticket offers from Broadway shows



Control de los
falsos
positivos

Es muy importante saber cuál es la tasa de falsos positivos, y si un resultado que ve es realmente "inusual".

Informar de todo lo que se ha probado, no sólo de los éxitos.

Comprobación de
hipótesis

Construir una hipótesis,
por ejemplo: "Los
gamers tienen redes
sociales menos activas
que los no gamers".

Comprobación de hipótesis

Definir un experimento para comprobar la hipótesis (modifique H si es necesario).

Elegir una población y un método de muestreo.

Crear una hipótesis nula H_0 .

Construir un estadístico de prueba.

Elegir un nivel de significación y el tamaño de la muestra.

Realizar el experimento.

Reportar todos los resultados.

Riesgos



Algunos aspectos importantes...



La privacidad de datos implica la gestión adecuada para minimizar el riesgo y proteger los datos confidenciales.



Muchos procesos de privacidad tradicionales no pueden manejar la escala y la velocidad requeridas.



Regulaciones de datos.



Cuanto más datos recopile, más importante será ser transparente con sus clientes sobre lo que está haciendo con sus datos, cómo los está almacenando.

Para la siguiente vez...



Sistema de almacenamiento distribuido

Referencias

- Jure Leskovec, Anand Rajaraman and Jeffrey D. Ullman. Mining of Massive Datasets. Second Edition. Cambridge University Press, 2014.
- Charu C. Aggarwal. Data Mining. Springer International Publishing, 2015.
- Jeffrey Vitter. Algorithms and Data Structures for External Memory. Now Foundations and Trends, 2008.