Fundamentos de Aprendizaje Automático

Máster Universitario en Ciencia de Datos aplicada a las Ciencias Sociales
por la Universidad de Granada y la Universidad de Salamanca

Métodos no supervisados

03. Reglas de asociación

UNIVERSIDAD
DE GRANADA

Índice

Sesión 4

- Reglas de asociación
  - Definición
  - Medidas
  - Algoritmos
  - Otros tipos de reglas de asociación
  - Aplicaciones

Sesión 5

- Uso de algoritmos de detección de reglas de asociación en Python
- Ejercicio práctico

1. Reglas de asociación

Motivación

> "The legend says that a study was done by a retail grocery store. The findings were that men between 30-40 years in age, shopping between 5pm and 7pm on Fridays, who purchased diapers were most likely to also have beer in their carts. This motivated the grocery store to move the beer aisle closer to the diaper aisle and wiz-boom-bang, an instant 35% increase in sales of both."
>
> https://canworksmart.com/diapers-beer-retail-predictive-analytics.

- En muchos casos se adjudica el análisis a un Walmart (no está claro)
- Ideas a considerar:
  - Restringir el análisis a un conjunto de tuplas (Ej: solo las transacciones en viernes, solo las transacciones en un rango horario…)
  - Considerar distinta granularidad en los datos antes de hacer el análisis

1. Reglas de asociación

Definición

Búsqueda de patrones de “asociación” en los datos

Es una técnica de aprendizaje/minería de datos no supervisado

No se conocen a priori la relación entre los datos, y si se conociesen, se busca que los datos las confirmen

Suelen representarse mediante “implicaciones” de la forma:

$$A \Rightarrow C$$

significando que en las transacciones donde se da el ítem A también se da C.

1. Reglas de asociación

Ejemplo

Ejemplo de cesta de la compra

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Buena parte de los ejemplos de esta sección están inspirados o tomados de: Pang-Ning Tan, Michael Steinbach & Vipin Kumar: Introduction to Data Mining Addison-Wesley, 2006. [capítulos 6&7].

1. Reglas de asociación

Ejemplo

Ejemplo de cesta de la compra:

- Regla de asociación

$$\{pañales\} \Rightarrow \{cerveza\}$$

- La regla implica co-ocurrencia.
- No implica causalidad.

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Buena parte de los ejemplos de esta sección están inspirados o tomados de: Pang-Ning Tan, Michael Steinbach & Vipin Kumar: Introduction to Data Mining Addison-Wesley, 2006. [capítulos 6&7].

1. Reglas de asociación

Nomenclatura

- Sea I un conjunto de ítems.
- T un conjunto de transacciones con ítems en I.

La regla de asociación

$$A \Rightarrow C, \qquad con \quad A, C \subseteq I, \qquad A, C \neq \emptyset \quad y \quad A \cap C = \emptyset$$

significa que cada transacción de T que contiene a *A*, contiene a *C*.

A y *C* son itemsets (conjuntos de ítems) no vacíos y disjuntos.

Al itemset *A* se le llama **antecedente** y a *C* **consecuente**.

Rakesh Agrawal, Tomasz Imielinski, Arun N. Swami: Mining Association Rules between Sets of Items in Large Databases. SIGMOD Conference 1993: 207-216

1. Reglas de asociación

Nomenclatura

- Para poder aplicar esta técnica, hay que identificar:
  - ¿Cuáles son los ítems?
  - ¿Cuáles son las transacciones?
- Por ejemplo, en la cesta de la compra:
  - Los ítems son los artículos de supermercado.
  - Cada cesta de la compra es una transacción.
  - La regla $\{pan, mantequilla\} \Rightarrow \{leche\}$ significa que en las cestas de la compra en las que hay pan y mantequilla, también hay leche.

1. Reglas de asociación

Ejemplo en BDR

- Ejemplo en bases de datos relacionales
  - Si tenemos una tabla con atributos y tuplas:
    - Cada pareja (atributo, valor) sería un ítem.
    - Cada tupla, entendida como conjunto de parejas (atributo, valor), sería una transacción.

ID	VIVIENDA	CALIDAD	SALARIO
1	APARTAMENTO	BAJA	MEDIO
2	PISO	MEDIA	MEDIO
3	DUPLEX	MEDIA	ALTO
4	UNIFAMILIAR	ALTA	ALTO

    - Transacciones:
      - (VIVIENDA, APARTAMENTO), (CALIDAD, BAJA), (SALARIO, MEDIO)
      - (VIVIENDA, PISO), (CALIDAD, MEDIA), (SALARIO, MEDIO)
      - (VIVIENDA, DUPLEX), (CALIDAD, MEDIA), (SALARIO, ALTO)
      - (VIVIENDA, UNIFAMILIAR), (CALIDAD, ALTA), (SALARIO, ALTO)

1. Reglas de asociación

Ejemplo en BDR

- Ejemplo en bases de datos relacionales

ID	VIVIENDA	CALIDAD	SALARIO
1	APARTAMENTO	BAJA	MEDIO
2	PISO	MEDIA	MEDIO
3	DUPLEX	MEDIA	ALTO
4	UNIFAMILIAR	ALTA	ALTO

  - Transacciones:
    - (VIVIENDA, APARTAMENTO), (CALIDAD, BAJA), (SALARIO, MEDIO)
    - (VIVIENDA, PISO), (CALIDAD, MEDIA), (SALARIO, MEDIO)
    - (VIVIENDA, DUPLEX), (CALIDAD, MEDIA), (SALARIO, ALTO)
    - (VIVIENDA, UNIFAMILIAR), (CALIDAD, ALTA), (SALARIO, ALTO)
  - Ejemplo de regla de asociación:

$$\{(CALIDAD, ALTA)\} \Rightarrow \{(SALARIO, ALTO)\}$$

1. Reglas de asociación

- Son de mucha utilidad, por ejemplo:
  - Si se identifica una regla del tipo: $\{producto\ 1\} \Rightarrow \{producto\ 2\}$

    - Puedo buscar estrategias para incrementar mis ventas:
      - Ponerlos juntos en la tienda.
      - Proponer promociones que incluyan al primer producto que podrían producir un incremento de ventas en el segundo.

    - Puedo hacerme una idea de qué productos podrían verse afectados si dejamos de vender el producto.

    - Puedo hacerme una idea de cómo debe ser el stock de los productos.

1. Reglas de asociación

Medidas

¿Cómo de buena es una regla?

- Hacen falta medidas que permitan distinguir reglas mejores de reglas peores:
  - La asociación descrita por la regla puede no cumplirse siempre:
    ¿con qué frecuencia lo hace?
    ¿sobre cuántos casos o ejemplos se sustenta?

- Hay una amplia variedad de propuestas.

- Las medidas convencionales o clásicas son el **soporte** y la **confianza**.

- Otras medidas: lift, convicción, factor de certeza…

1. Reglas de asociación

Medidas

- Soporte
  - Si consideramos un conjunto de ítems o itemset.
    - El **soporte de un itemset** es la frecuencia con la que ocurre en la base de datos. X ocurre en una transacción t, si $X \subseteq t$.

$$Soporte\ (X) = \frac{Nº\ de\ ocurrencias\ de\ X}{Nº\ total\ de\ transacciones\ en\ la\ BD}$$

  - Si consideramos una regla
    - El **soporte de una regla** es el número de transacciones que contienen a la unión de los itemsets de la regla.

$$Soporte\ (A \Rightarrow C) = Soporte\ (A \cup C) = \frac{Nº\ de\ ocurrencias\ de\ A \cup C}{Nº\ total\ de\ transacciones\ en\ la\ BD}$$

1. Reglas de asociación

Ejemplo en BDR

- Ejemplo en bases de datos relacionales

ID	VIVIENDA	CALIDAD	SALARIO
1	APARTAMENTO	BAJA	MEDIO
2	PISO	MEDIA	MEDIO
3	DUPLEX	MEDIA	ALTO
4	UNIFAMILIAR	ALTA	ALTO

- Transacciones:
  - (VIVIENDA, APARTAMENTO), (CALIDAD, BAJA), (SALARIO, MEDIO)
  - (VIVIENDA, PISO), (CALIDAD, MEDIA), (SALARIO, MEDIO)
  - (VIVIENDA, DUPLEX), (CALIDAD, MEDIA), (SALARIO, ALTO)
  - (VIVIENDA, UNIFAMILIAR), (CALIDAD, ALTA), (SALARIO, ALTO)

- Regla de asociación: $\{(CALIDAD, ALTA)\} \Rightarrow \{(SALARIO, ALTO)\}$
- Soporte del ítem $(CALIDAD, ALTA)$:

$$Soporte\left((CALIDAD, ALTA)\right) = \frac{Nº\ de\ ocurrencias\ de(CALIDAD, ALTA)}{Nº\ total\ de\ transacciones\ en\ la\ BD} = \frac{1}{4}$$

1. Reglas de asociación

Soporte

El soporte toma valores entre 0 y 1.

- 0 -> No se da en la BD.
- 1 -> Está presente **en todas** las transacciones.

• Ejemplo:

ID	VIVIENDA	CALIDAD	SALARIO
1	APARTAMENTO	BAJA	MEDIO
2	PISO	MEDIA	MEDIO
3	DUPLEX	MEDIA	ALTO
4	UNIFAMILIAR	ALTA	ALTO

$$Soporte\ (\{(Vivienda, Apartamento)\}) = \frac{4}{4} = 1$$

$$Soporte\ (\{(Calidad, Alta)\}) = \frac{0}{4} = 0$$

1. Reglas de asociación

Medidas

- Confianza
  - Si consideramos una regla $A \Rightarrow C$:

$$Confianza\ (A \Rightarrow C) = \frac{Soporte\ (A \cup C)}{Soporte(A)}$$

  - Puede verse como una probabilidad condicionada:

$$prob(C|A) = \frac{prob(A \wedge C)}{prob(A)}$$

1. Reglas de asociación

Ejemplo en BDR

ID	VIVIENDA	CALIDAD	SALARIO
1	APARTAMENTO	BAJA	MEDIO
2	PISO	MEDIA	MEDIO
3	DUPLEX	MEDIA	ALTO
4	UNIFAMILIAR	ALTA	ALTO

- Transacciones:
  - (VIVIENDA, APARTAMENTO), (CALIDAD, BAJA), (SALARIO, MEDIO)
  - (VIVIENDA, PISO), (CALIDAD, MEDIA), (SALARIO, MEDIO)
  - (VIVIENDA, DUPLEX), (CALIDAD, MEDIA), (SALARIO, ALTO)
  - (VIVIENDA, UNIFAMILIAR), (CALIDAD, ALTA), (SALARIO, ALTO)

- Regla de asociación: $\{(CALIDAD, ALTA)\} \Rightarrow \{(SALARIO, ALTO)\}$

$$Soporte\ (\{(CALIDAD, ALTA)\}) = 1/4$$

$$Soporte\ (\{(CALIDAD, ALTA) \cup (SALARIO, ALTO)\}) = 1/4$$

$$Confianza\ \big((CALIDAD, ALTA) \Rightarrow (SALARIO, ALTO)\big) = \frac{Sop(\{(CALIDAD,\ ALTA) \cup (SALARIO,\ ALTO)\})}{Sop(\{(CALIDAD,\ ALTA)\})} = 1$$

1. Reglas de asociación

Ejemplo en BD Transaccional

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

$$Soporte(\{cerveza\}) = 4/5 = \ 0.8$$

$$Soporte(\{pañales\}) = 3/5 = \ 0.6$$

$$Soporte(\{cerveza\} \Rightarrow \{pañales\}) =$$

$$= Soporte(\{pañales, cerveza\}) = 3/5$$

$$= 0.6 = 60\%$$

$$Confianza(\{cerveza\} \Rightarrow \{pañales\}) =$$

$$= \frac{Soporte(\{pañales, cerveza\})}{Soporte(\{cerveza\})} = \frac{3/5}{4/5} = \frac{3}{4} = 0.75 = 75\%$$

1. Reglas de asociación

Ejemplo en BD Transaccional

$$Soporte(\{cerveza\}) = 4/5 = \ 0.8$$

$$Soporte(\{pañales\}) = 3/5 = \ 0.6$$

$$Soporte(\{cerveza\} \Rightarrow \{pañales\}) =$$

$$= Soporte(\{pañales, cerveza\}) = 3/5$$

$$= 0.6 = 60\%$$

$$Confianza(\{pañales\} \Rightarrow \{cerveza\}) =$$

$$= \frac{Soporte(\{pañales, cerveza\})}{Soporte(\{pañales\})} = \frac{3/5}{3/5} = 1 = 100\%$$

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

1. Reglas de asociación

Confianza

La confianza toma valores entre 0 y 1.

- En una regla del tipo $A \Rightarrow C$:
  - 0 -> Nunca ocurre C, cuando ocurre A.
  - 1 -> Siempre que ocurre A, ocurre C.
- Ejemplo:

ID	VIVIENDA	CALIDAD	SALARIO
1	APARTAMENTO	BAJA	MEDIO
2	PISO	MEDIA	MEDIO
3	DUPLEX	MEDIA	ALTO
4	UNIFAMILIAR	ALTA	ALTO

$$Confianza\big((VIVIENDA, PISO) \Rightarrow (SALARIO, ALTO)\big) = \frac{0}{1/4} = 0$$

$$Confianza\left((CALIDAD, ALTA) \Rightarrow (SALARIO, ALTO)\right) = \frac{1/4}{1/4} = 1$$

1. Reglas de asociación

Otras medidas

- **Lift** (Elevación):

Rango: $[0, \infty)$

Definición: Es la razón de la confianza de la regla $X \Rightarrow Y$ a la proporción de transacciones que contienen Y. Indica cuánto más probable es encontrar Y en las transacciones que contienen X en comparación con una situación aleatoria.

Interpretación:

- Lift = 1: No hay ninguna relación entre el antecedente y el consecuente.
- Lift > 1: Existe una relación positiva entre el antecedente y el consecuente.
- Lift < 1: Existe una relación negativa entre el antecedente y el consecuente.

1. Reglas de asociación

Otras medidas

- **Leverage** (Apalancamiento):

Definición: Mide la diferencia entre el soporte observado de X U Y y el soporte esperado si X e Y fueran independientes.

Rango: [-1, 1]

Interpretación:

- Leverage = 0: No hay ninguna relación entre el antecedente y el consecuente.
- Leverage > 0: Relación positiva entre el antecedente y el consecuente.
- Leverage < 0: Relación negativa entre el antecedente y el consecuente

1. Reglas de asociación

Otras medidas

- **Conviction** (Convicción):

Rango: [0, ∞)

Definición: Mide la dependencia de la regla X ⇒ Y. Es la razón de la proporción esperada de transacciones que contienen X y no Y a la proporción observada de esas transacciones. Una convicción más alta indica una regla más fuerte.

Interpretación:

• Conviction = 1: No hay ninguna relación entre el antecedente y el consecuente.

• Conviction > 1: Cuanto mayor es el valor, mayor es la certeza de que el antecedente implica el consecuente.

1. Reglas de asociación

Otras medidas

- **Métrica de Zhang:**

Rango: [-1, 1]

Se basa en la idea de medir el cambio relativo en la probabilidad condicional de encontrar el ítem consecuente dado el ítem antecedente, en comparación con encontrar el ítem consecuente sin ninguna condición.

Interpretación:

•**Valor positivo**: Un valor positivo de la métrica de Zhang indica que el ítem consecuente Y es más probable de aparecer cuando el ítem antecedente X está presente.

•**Valor negativo**: Un valor negativo sugiere que el ítem consecuente Y es menos probable de aparecer cuando el ítem antecedente X está presente.

•**Valor cercano a cero**: Indica que no hay una relación fuerte entre los ítems X e Y.

Problema de extracción

Dado un conjunto de transacciones T, encontrar todas las reglas de asociación en función de ciertos requisitos de calidad:

- Soporte mayor o igual que un cierto umbral (**MinSop**).
- Confianza mayor o igual que un cierto umbral (**MinConf**).

1. Reglas de asociación

Algoritmo

Solución por fuerza bruta

- Enumerar todas las reglas de asociación posibles.
- Calcular el soporte y la confianza de cada regla.
- Eliminar las reglas que no superen los umbrales de soporte y confianza (MinSop y MinConf).

- Computacionalmente prohibitivo…

1. Reglas de asociación

Algoritmo

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Ejemplo:

Reglas derivadas del itemset {pan, pañales, cerveza}

{pan} ⇒ {pañales, cerveza}, Sop=0.4, Conf=2/4=0.5

{pañales} ⇒ {pan, cerveza}, Sop=0.4, Conf=2/3=0.67

{cerveza} ⇒ {pan, pañales}, Sop=0.4, Conf=2/4=0.5

{pan, pañales} ⇒ {cerveza}, Sop=0.4, Conf=2/2=1

{pan, cerveza} ⇒ {pañales}, Sop=0.4, Conf=2/3=0.67

{pañales, cerveza} ⇒ {pan}, Sop=0.4, Conf=2/3=0.67

1. Reglas de asociación

Algoritmo

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Ejemplo:

Reglas derivadas de {pan, pañales, cerveza}

- Observaciones
  - Todas las reglas anteriores se obtienen dividiendo en dos partes el mismo itemset ({pan, pañales, cerveza}).
  - **¡Tienen el mismo soporte!** aunque su confianza pueda variar.

1. Reglas de asociación

Algoritmo

- Solución en dos etapas
  - **Generación de itemsets frecuentes:**

    Identificar los itemsets con soporte por encima de un umbral fijado por el usuario (MinSop).

  - **Generación de reglas de asociación:**

    Obtener reglas de asociación con una confianza elevada a partir de cada itemset frecuente (cada regla es una partición binaria del itemset). También se fija un umbral de confianza (MinConf).

    Nota: La generación de itemsets frecuentes sigue siendo computacionalmente costosa.

1. Reglas de asociación

Algoritmos

- Estrategias
  - Reducir el número de candidatos
    - Tratando de podar (no estudiar todos los candidatos posibles).
    - Ejemplo: Algoritmos Apriori y DHP [Direct Hashing and Pruning]
  - Reducir el número de transacciones, conforme aumenta el tamaño del itemset.
    - Ejemplo: Algoritmo AprioriTID
  - Reducir el número de comparaciones
    - Mediante el uso de estructuras de datos eficientes.
    - Ejemplo: Algoritmo FP-Growth

1. Reglas de asociación

Algoritmos

Estrategias

- Reducir el número de candidatos
  - Tratando de podar (no estudiar todos los candidatos posibles).
  - Ejemplo: Algoritmos Apriori y DHP [Direct Hashing and Pruning]

- Reducir el número de transacciones, conforme aumenta el tamaño del itemset.
  - Ejemplo: Algoritmo AprioriTID

- Reducir el número de comparaciones
  - Mediante el uso de estructuras de datos eficientes.
  - Ejemplo: Algoritmo FP-Growth

1. Reglas de asociación

Algoritmos

Reducción del número de candidatos (**propiedad Apriori**)

- Si un itemset es frecuente, también lo son todos sus subconjuntos.

- **¿Por qué?** Porque el soporte de un itemset nunca puede ser mayor que el de cualquiera de sus subconjuntos:

$$\forall X,Y,\ X \subseteq Y \Rightarrow Soporte(X) \geq Soporte(Y)$$

Formalmente, esta propiedad se conoce con el nombre de **anti-monotonía del soporte**.

1. Reglas de asociación

Algoritmos

Reducción del número de candidatos (propiedad Apriori)

$$\forall X, Y, \qquad X \subseteq Y \Rightarrow Soporte(X) \geq Soporte(Y)$$

Ejemplo:

MinConf=0.7

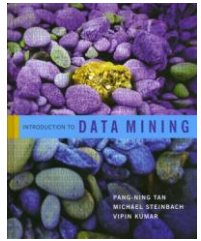
ID	VIVIENDA	CALIDAD	SALARIO
1	APARTAMENTO	BAJA	MEDIO
2	PISO	MEDIA	MEDIO
3	DUPLEX	MEDIA	ALTO
4	UNIFAMILIAR	ALTA	ALTO

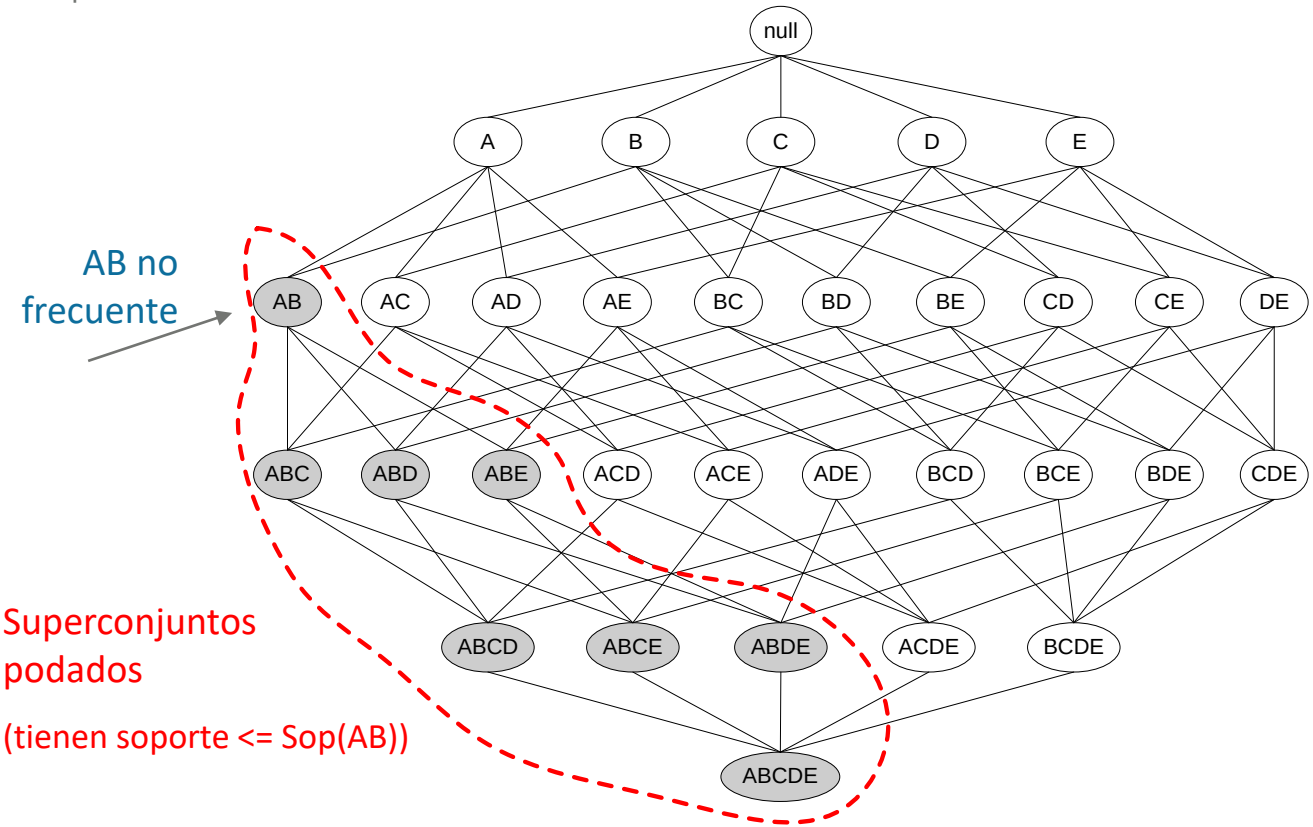
$$Soporte\ (\{(CALIDAD, MEDIA)\}) = \frac{2}{4} = 0.5$$

$$Soporte\ (\{(VIVIENDA\ PISO), (CALIDAD, MEDIA)\}) = \frac{1}{4} = 0.25$$

1. Reglas de asociación

Propiedad Apriori





1. Reglas de asociación

Algoritmo Apriori

Tablas

L[k] = Conjunto de k-itemsets frecuentes

C[k] = Conjunto de k-itemsets potencialmente frecuentes (candidatos)

Algoritmo

Generar L[1] (patrones frecuentes de tamaño 1, i.e. ítems)

Repetir mientras se descubran nuevos itemsets frecuentes:

a) Generar los candidatos C[k+1] a partir de los patrones frecuentes L[k].

b) Contabilizar el soporte de cada candidato de C[k+1] recorriendo la base de datos secuencialmente.

c) Eliminar candidatos no frecuentes, dejando en L[k+1] solo aquellos que son frecuentes.

1. Reglas de asociación

Algoritmo Apriori

TID	Artículos
1	Pan, leche
2	Pan, pañales, cerveza, huevos
3	Leche, pañales, cerveza, vino
4	Pan, leche, pañales, cerveza
5	Pan, leche, pañales, vino

$$Mínimo\ Soporte = \frac{3}{5}$$

$$Soporte\ (pan) = \frac{4}{5}$$

$$Soporte\ (leche) = \frac{4}{5}$$

$$Soporte\ (pañales) = \frac{4}{5}$$

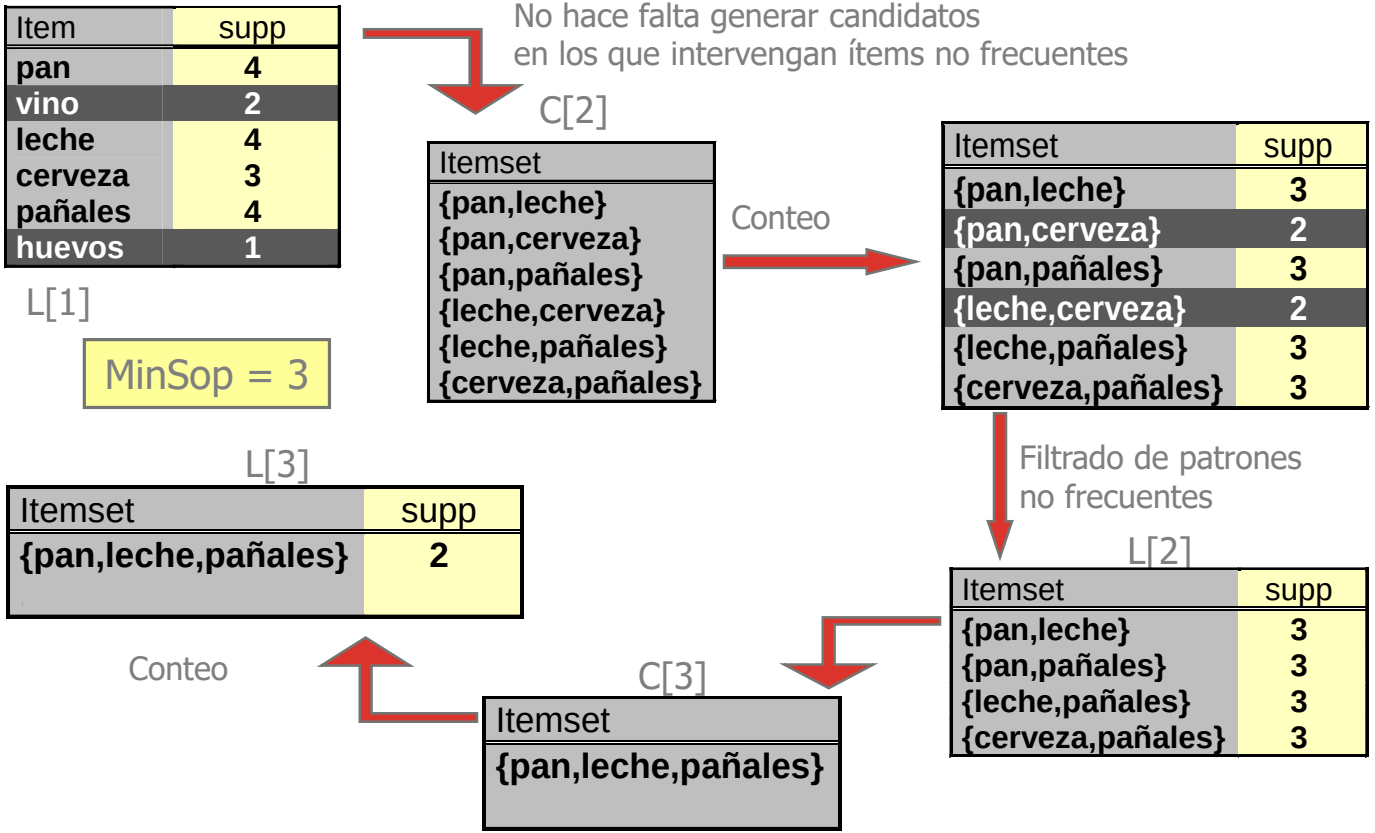
$$Soporte\ (cerveza) = \frac{3}{5}$$

$$Soporte\ (huevos) = \frac{1}{5}$$

$$Soporte\ (vino) = \frac{2}{5}$$

1. Reglas de asociación

Algoritmo Apriori



1. Reglas de asociación

Algoritmo Apriori

Generación de reglas

- Dado un itemset frecuente L, se encuentran todos los subconjuntos no vacíos $f \subset L$ tales que la regla $\{f \Rightarrow L - f\}$ satisfaga el umbral de confianza mínima (MinConf).

Ejemplo

- A partir del itemset frecuente {A,B,C,D}, se generan las siguientes reglas candidatas:

$ABC \Rightarrow D$, $ABD \Rightarrow C$, $ACD \Rightarrow B$, $BCD \Rightarrow A$,
$A \Rightarrow BCD$, $B \Rightarrow ACD$, $C \Rightarrow ABD$, $D \Rightarrow ABC$
$AB \Rightarrow CD$, $AC \Rightarrow BD$, $AD \Rightarrow BC$, $BC \Rightarrow AD$, $BD \Rightarrow AC$, $CD \Rightarrow AB$

- Si $|L| = k$, entonces hay $2^k - 2$ reglas de asociación candidatas (ignorando $L \Rightarrow \varnothing$ y $\varnothing \Rightarrow L$)

1. Reglas de asociación

Algoritmo Apriori

¿Cómo generar las reglas de forma eficiente?

- ¿Es la confianza anti-monótona como el soporte?

  NO: La confianza de $ABC \rightarrow D$ puede ser mayor o menor que la confianza de $AB \rightarrow D$.

- Pero la confianza de las reglas generadas de un mismo itemset tienen una propiedad anti-monótona:

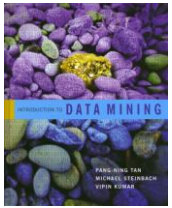
  Por ejemplo: $L = \{A,B,C,D\}$

  $Conf(ABC \Rightarrow D) \geq Conf(AB \Rightarrow CD) \geq Conf(A \Rightarrow BCD)$

- La confianza es **anti-monótona con respecto al número de ítems en el consecuente** (la parte derecha de la regla).

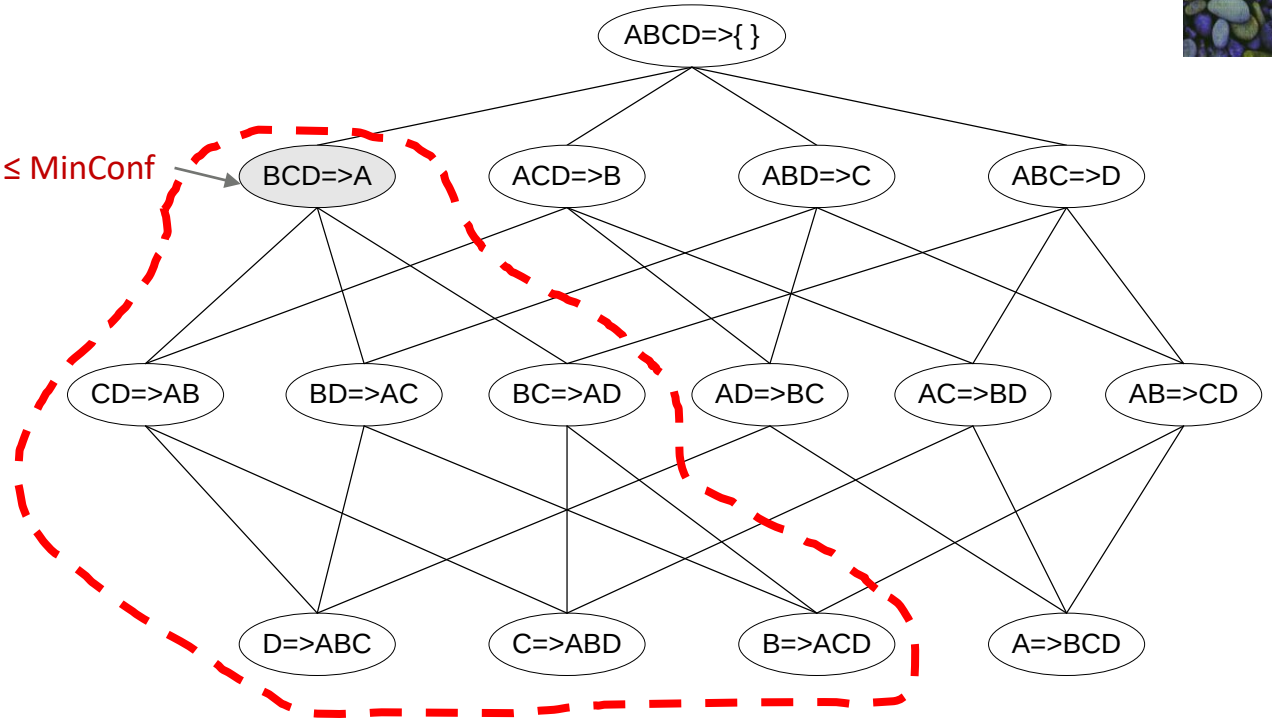
1. Reglas de asociación

Algoritmo Apriori



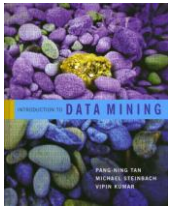
¿Cómo generar las reglas de forma eficiente?

Si Conf(BCD⇒A) ≤ MinConf



1. Reglas de asociación

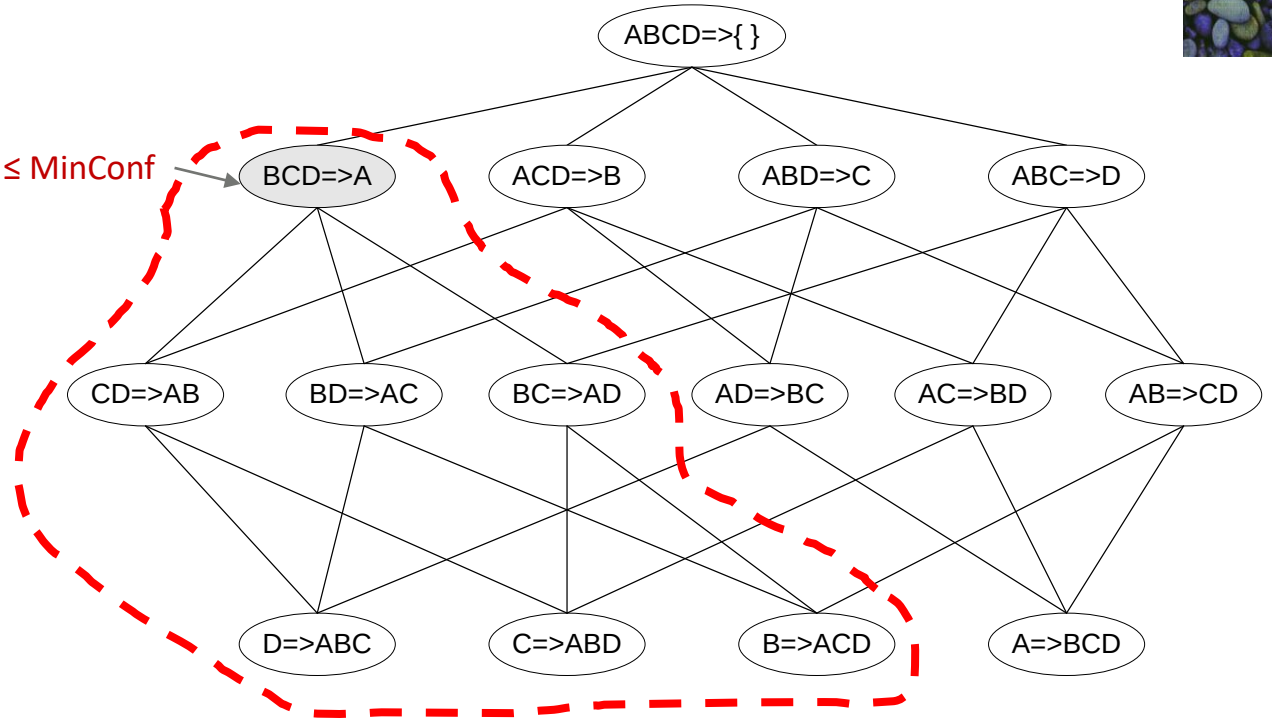
Algoritmo Apriori



¿Cómo generar las reglas de forma eficiente?

Si Conf(BCD⇒A) ≤ MinConf

Todas las reglas con A en el consequente se podan



1. Reglas de asociación

Ejercicio

- Proponga un conjunto sencillo de transaciones (al menos 6) considerando el tipo de datos presente en el cubo.

- Aplique el algoritmo Apriori para obtener algunas reglas de asociación en base a los umbrales de
  - Mínimo soporte (por ejemplo 0.2)
  - Mínima confianza (por ejemplo 0.7)

1. Reglas de asociación

Problemas en la obtención de las reglas

- Soporte o confianza demasiado bajos
  - Puede provocar que haya demasiadas reglas -> más difícil de inspeccionar
  - El algoritmo se ralentiza -> Explosión combinatoria

- Soporte o confianza demasiado altos
  - Puede provocar que no haya reglas

Además…

Podemos elegir otras medidas distintas de la confianza: Lift, convicción, etc.

1. Reglas de asociación

¿Atributos numéricos continuos?

- Puede que no se repita justamente la misma cantidad -> Poca probabilidad de alcanzar el umbral MinSop
- **Solución:** Discretizar en intervalos
- **Problema:** Alta dependencia de la elección de intervalos

SALARIO:
{[0,30k),[30k,60k),[60k,)}
{[0,31k),[31k,62k),[62k,)}

ID	SALARIO
1	25000
2	30000
3	61000
4	72000

ID	SALARIO
1	[0,30k)
2	[30k,60k)
3	[60k,)
4	[60k,)

ID	SALARIO
1	[0,31k)
2	[0,31k)
3	[31k,62k)
4	[62k,)

- Discretizaciones parecidas pueden dar lugar a conjuntos de reglas distintas

1. Reglas de asociación

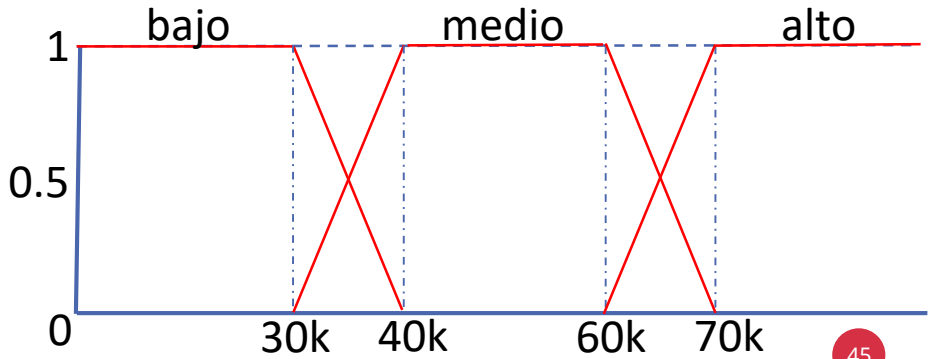
Otros tipos de reglas de asociación

¿Atributos numéricos continuos?

- A tener en cuenta:
  - Estudio de la distribución de los datos antes de discretizar
  - Uso de “bordes suaves” -> conjuntos difusos (con etiquetas lingüísticas asociadas)

Solución -> **Reglas de asociación difusas**

ID	VIVIENDA	CALIDAD	SALARIO
1	APARTAMENTO	BAJA, 0.2	BAJO, 0.8
2	PISO	MEDIA, 0.5	MEDIO, 1
3	DUPLEX	MEDIA, 1	ALTO, 0.3
4	UNIFAMILIAR	ALTA, 0.8	ALTO, 0.8



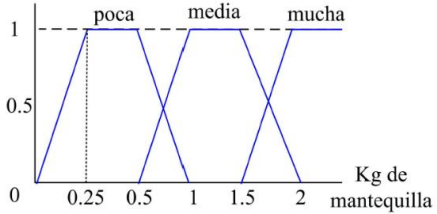
1. Reglas de asociación

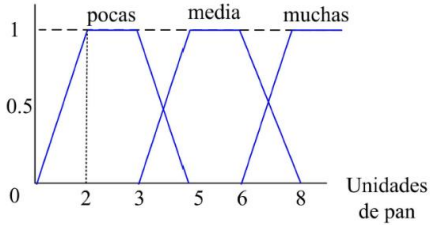
Otros tipos de reglas de asociación

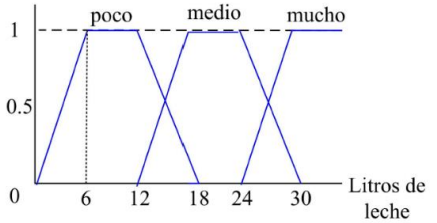
Reglas de asociación difusas

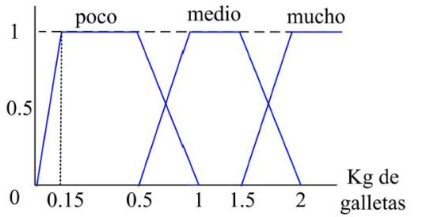
D_2	mantequilla (kg)	pan (un.)	leche (l)	galletas (kg)
B_1	1	1	13	1.75
B_2	1.25	2	18	2
B_3	0.25	1	5	1
B_4	0.25	5	1	0.5
B_5	0.1	2	0	0.5
B_6	0.5	0	5	1

$$\{\text{poca leche}\} \Rightarrow \{\text{poca mantequilla}\}$$







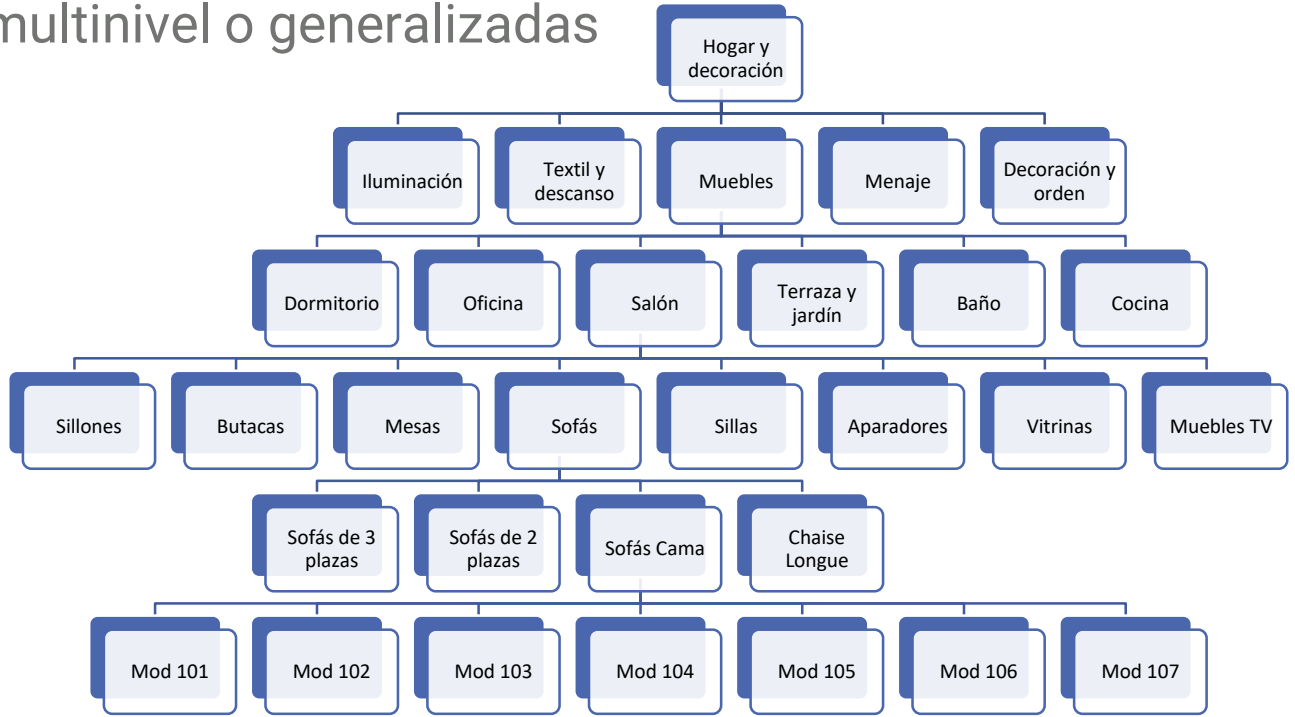


M. Delgado et al. Pattern Extraction from Bag Databases. *Int. J. Uncertain. Fuzziness Knowl. Based Syst*. 16 (4): 475-494 (2008)

1. Reglas de asociación

Otros tipos de reglas de asociación

Reglas multinivel o generalizadas



1. Reglas de asociación

Otros tipos de reglas de asociación

Reglas multinivel o generalizadas

- ¿Por qué utilizar jerarquías de conceptos?
  - Porque las reglas que involucran artículos en los niveles más bajos puede que no tengan soporte suficiente como para aparecer en algún patrón frecuente.
  - Porque las reglas a niveles bajos de la jerarquía pueden ser demasiado específicas.

    mod 104 ⇒ mod 207,

    mod 106 ⇒ mod 204,

    …

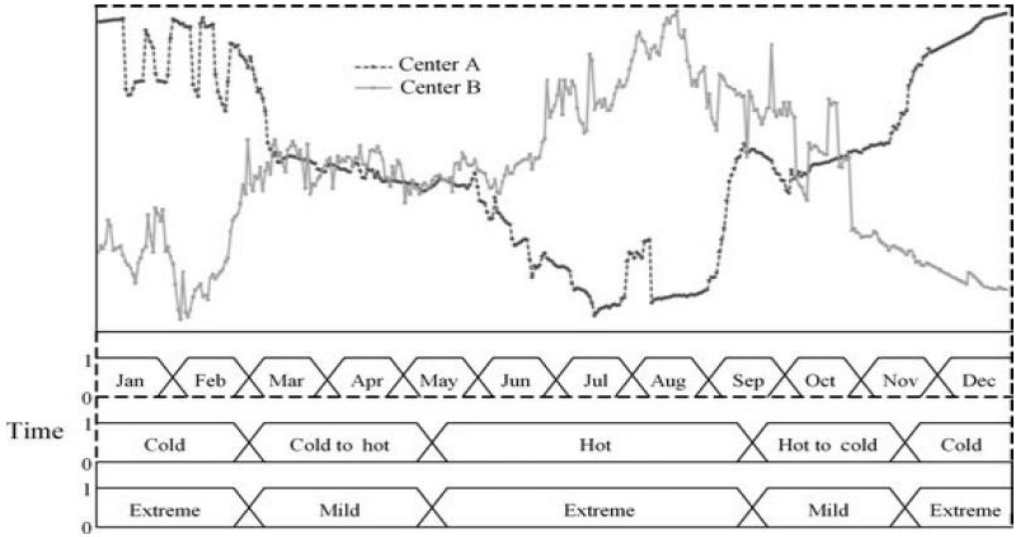
    indican una asociación entre sofás-cama y mesita-bajera.
- Por ejemplo: mínimo soporte = 0.25
  - leche sin lactosa ⇒ pan de centeno Soporte = 0.2
  - leche especial ⇒ pan especial Soporte = 0.3
- Por el contrario, si el soporte es excesivo, podemos bajar en la jerarquía (aumentar granularidad).

1. Reglas de asociación

Otros tipos de reglas de asociación

Reglas temporales

Uso combinado de etiquetas difusas y restricción en distintos niveles de tiempo



Most of days of December, both series exhibit a local change with similar variation but different sign (0.984944).

R. Castillo-Ortega et al. Linguistic query answering on data cubes with time dimension. *Int. J. Intell. Syst.* 26(10) 1002-1021 (2011)

1. Reglas de asociación

Otros tipos de reglas de asociación

Reglas de excepción

- Ausencia de ítems: negación
  - Excepciones

$$X \Rightarrow Y \quad \text{(alto soporte y confianza)}$$
$$\{X, E\} \Rightarrow \neg Y \text{ (alta confianza)}$$

antibióticos ⇒ recuperación

{antibióticos, estafilococos} ⇒ ¬ recuperación

con la ayuda de *antibióticos*, el paciente normalmente tiende a *recuperarse*, a menos que aparezcan *estafilococos*

M. Delgado et al. New Approaches for Discovering Exception and Anomalous Rules. *Int. J. Uncertain. Fuzziness Knowl. Based Syst*. 19 (2): 361-399 (2011)

1. Reglas de asociación

Otros tipos de reglas de asociación

Reglas anómalas

- Ausencia de ítems: negación
  - Anomalías

$$X \Rightarrow Y \text{ (alto soporte y confianza)}$$
$$\{X, \neg Y\} \Rightarrow A \text{ (alta confianza)}$$

$$\text{síntomas} \Rightarrow \text{enfermedad 1}$$
$$\{\text{síntomas}, \neg \text{ enfermedad 1}\} \Rightarrow \text{enfermedad 2}$$

Normalmente con ciertos síntomas, el paciente suele tener la enfermedad 1, o bien la enfermedad 2 (menos común)

M. Delgado et al. New Approaches for Discovering Exception and Anomalous Rules. Int. J. Uncertain. Fuzziness Knowl. Based Syst. 19 (2): 361-399 (2011)

1. Reglas de asociación

Otros tipos de reglas de asociación

Dependencias graduales

D_2	mantequilla (kg)	pan (un.)	leche (l)	galletas (kg)
B_1	1	1	13	1.75
B_2	1.25	2	18	2
B_3	0.25	1	5	1
B_4	0.25	5	1	0.5
B_5	0.1	2	0	0.5
B_6	0.5	0	5	1

$(>, \text{pan}) \Rightarrow (>, \text{mantequilla})$

cuanto mayor (menor) es la cantidad de pan, mayor (menor) es la cantidad de mantequilla

M. Delgado et al. Pattern Extraction from Bag Databases. *Int. J. Uncertain. Fuzziness Knowl. Based Syst*. 16 (4): 475-494 (2008)

1. Reglas de asociación

Otros tipos de reglas de asociación

- Reglas multinivel o generalizadas
- Reglas de asociación difusas
- Reglas temporales
- Reglas de excepción
- Reglas anómalas
- Dependencias graduales
- …

1. Reglas de asociación

Ejercicio

- ¿Qué tipo de reglas se obtendrían en la siguiente BD? Pon un ejemplo

Población (ID)	Contaminación	Precipitaciones	Temperatura media
Granada	Alta	Media	Alta
Sevilla	Muy alta	Poca	Muy alta
Málaga	Alta	Baja	Media
…			

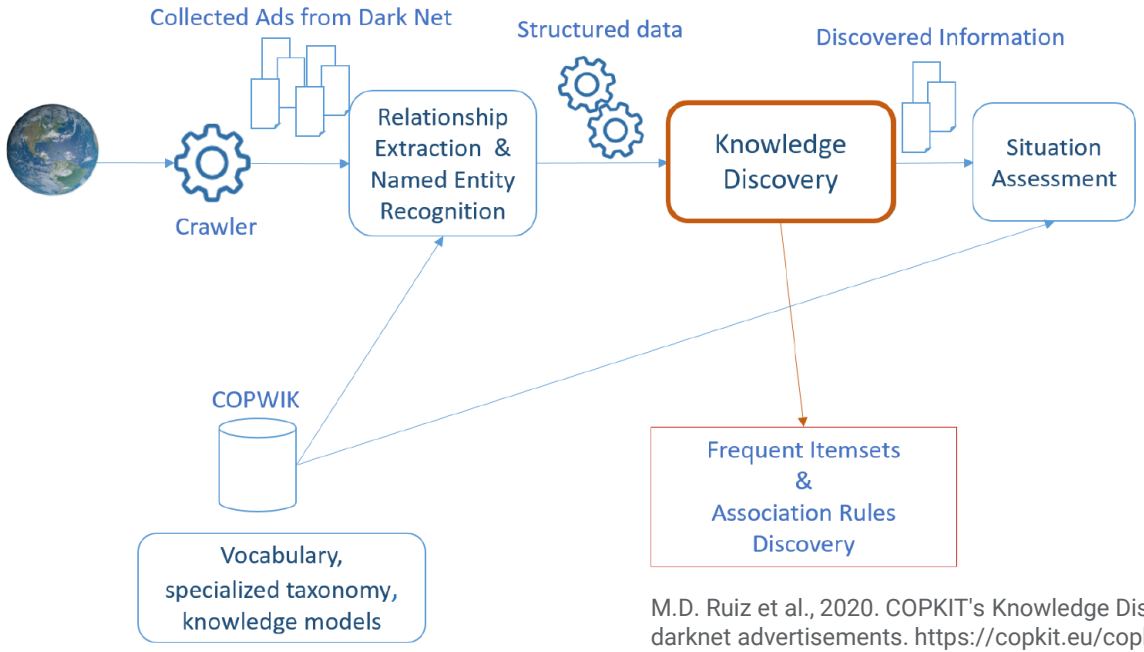
- ¿Y si las transacciones cambian su Id? ¿Cambia el significado de las reglas obtenidas?

Comarca (ID)	Contaminación	Precipitaciones	Temperatura media
Alpujarra Granadina	Alta	Media	Alta
Alhama	Muy alta	Poca	Muy alta
Baza	Alta	Baja	Media
…			

1. Reglas de asociación

Aplicaciones

Análisis de anuncios en Darknet:



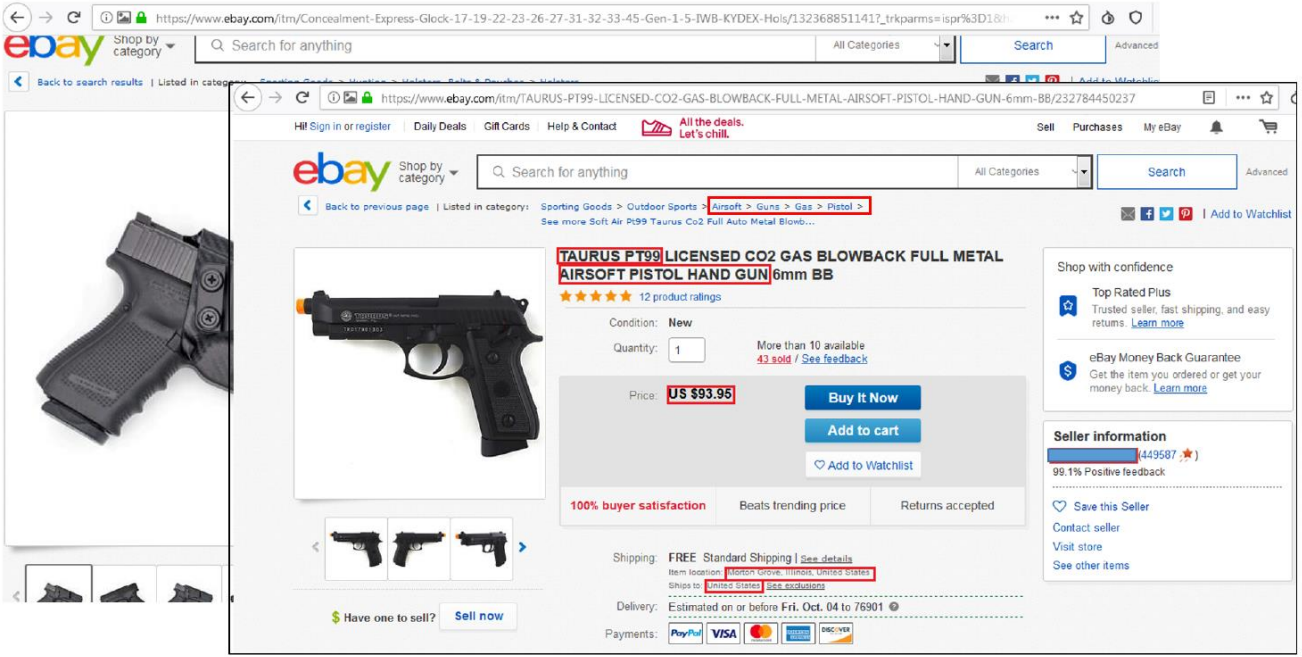
M.D. Ruiz et al., 2020. COPKIT's Knowledge Discovery tool: finding patterns in darknet advertisements. https://copkit.eu/copkits-knowledge-discovery-tool-finding-patternsin-darknet-advertisements/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687

1. Reglas de asociación

Aplicaciones

Análisis de anuncios en Darknet:



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687

1. Reglas de asociación

Aplicaciones

Análisis de anuncios en Darknet:

A	B	C	D	E	F	G	H	I	J	K
Market	Vendor	SoldItem	ShipFrom	Keywords	ArmType	Datetime	Category	Month	DayWeek	Daytime
Market_Abraxa	Vendor_user1	SoldItem_ESC	ShipFrom_W	Keywords_('p	ArmType_	Lir Datetime_07	Category_pistol	Month_Jun	DayWeek_Sunday	day
Market_Abraxa	Vendor_user2	SoldItem_Cro	ShipFrom_Ge	Keywords_('p	ArmType_	Cr Datetime_03	Category_pistol	Month_Jul	DayWeek_Friday	day
Market_Abraxa	Vendor_user3	SoldItem_Tau	ShipFrom_Ge	Keywords_('p	ArmType_	Ha Datetime_03	Category_pistol	Month_Aug	DayWeek_Monday	day
Market_Abraxa	Vendor_user3	SoldItem_Glo	ShipFrom_Ge	Keywords_('g	ArmType_	Ha Datetime_03	Category_glock 21	Month_Aug	DayWeek_Monday	day
Market_Abraxa	Vendor_user3	SoldItem_Col	ShipFrom_Ur	Keywords_('a	ArmType_	Ha Datetime_05	Category_ar 15	Month_Sep	DayWeek_Saturday	day
Market_Abraxa	Vendor_user3	SoldItem_Ru	ShipFrom_Ur	Keywords_('s	ArmType_	Ha Datetime_05	Category_sr 1911	Month_Sep	DayWeek_Saturday	day
Market_Abraxa	Vendor_user3	SoldItem_Col	ShipFrom_Ur	Keywords_('a	ArmType_	Ha Datetime_05	Category_ar 15	Month_Sep	DayWeek_Saturday	day
Market_Abraxa	Vendor_user3	SoldItem_Glo	ShipFrom_Ur	Keywords_('g	ArmType_	Ha Datetime_18	Category_glock 19	Month_Sep	DayWeek_Friday	night
Market_Abraxa	Vendor_user4	SoldItem_Glo	ShipFrom_Au	Keywords_('g	ArmType_	Ha Datetime_26	Category_glock	Month_Oct	DayWeek_Monday	night
Market_Agora	Vendor_user5	SoldItem_Spr	ShipFrom_Ur	Keywords_('s	ArmType_	Ha Datetime_08	Category_springfiel	Month_Jul	DayWeek_Wednesday	night
Market_Agora	Vendor_user5	SoldItem_Ber	ShipFrom_Ur	Keywords_('b	ArmType_	Ha Datetime_08	Category_beretta	Month_Jul	DayWeek_Wednesday	night
Market_Alphab	Vendor_user6	SoldItem_Glo	ShipFrom_W	Keywords_('g	ArmType_	Ha Datetime_23	Category_glock 26	Month_May	DayWeek_Saturday	night
Market_Alphab	Vendor_user7	SoldItem_VE	ShipFrom_W	Keywords_('a	ArmType_	To Datetime_16	Category_ammuniti	Month_Jun	DayWeek_Tuesday	day
Market_Alphab	Vendor_user8	SoldItem_Glo	ShipFrom_W	Keywords_('g	ArmType_	Ha Datetime_12	Category_glock 26	Month_Jun	DayWeek_Friday	night
Market_Alphab	Vendor_user8	SoldItem_Ma	ShipFrom_W	Keywords_('p	ArmType_	Ha Datetime_15	Category_p64	Month_Jun	DayWeek_Monday	day

$$\text{Vendor}=\text{user15} \Rightarrow \text{Category}=\text{Ammunition}, \text{ShipFrom}=\text{Worldwide}\ (0.08, 0.96)$$

La mayoría de los anuncios con el vendedor “user15” ofrecían munición que podía ser mandada desde cualquier sitio (96% de confianza), soporte 8% de anuncios

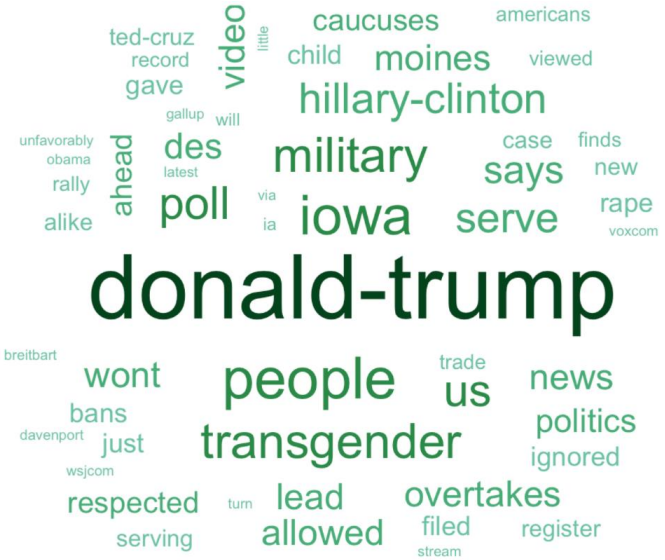
1. Reglas de asociación

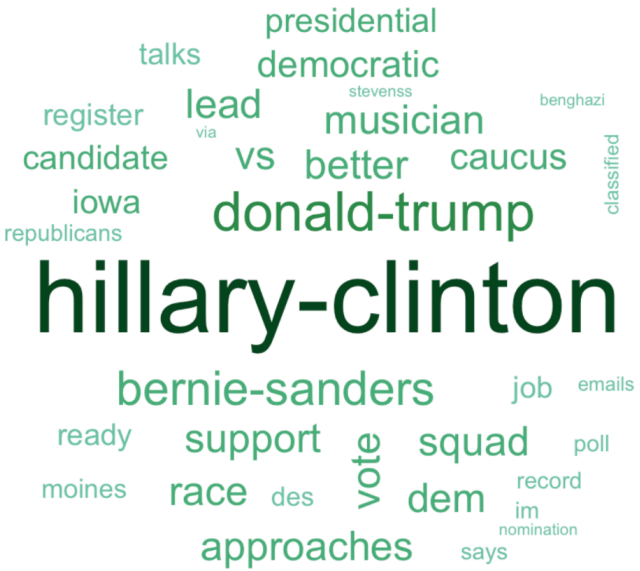
Aplicaciones

Análisis de sentimientos en Twitter

>1.6 Millones de tweets sobre las elecciones americanas de 2016

Itemsets frecuentes después del procesamiento:



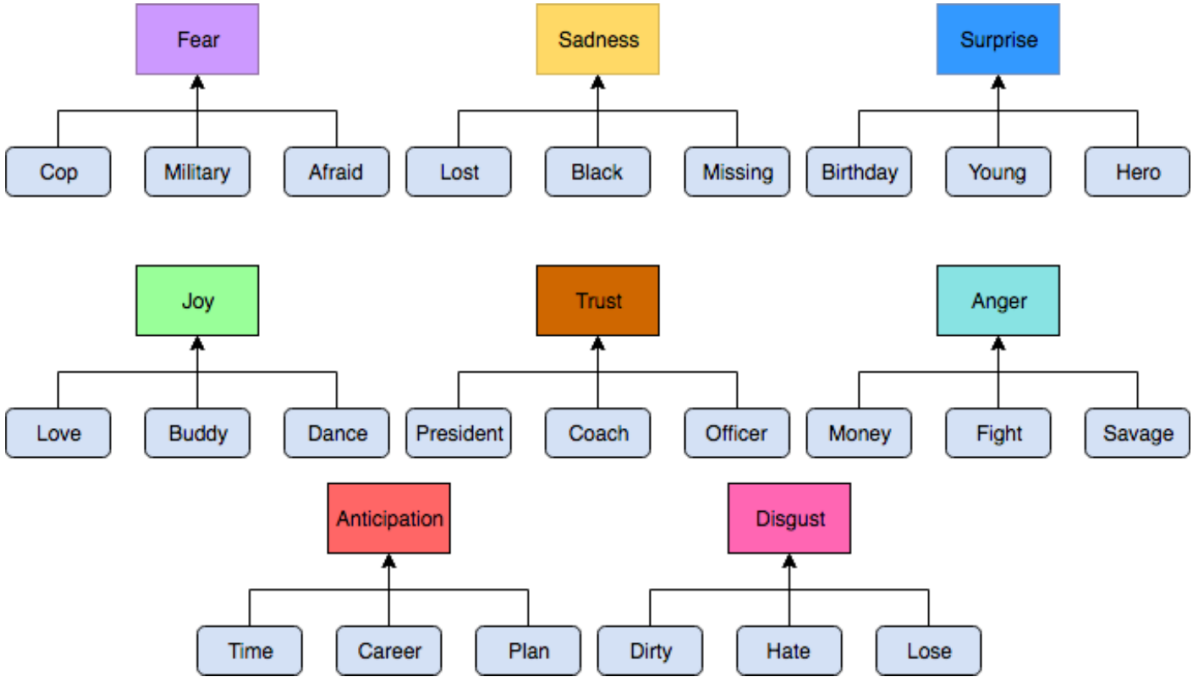


1. Reglas de asociación

Aplicaciones

Análisis de sentimientos en Twitter

Generalización usando sentimientos:



1. Reglas de asociación

Aplicaciones

Análisis de sentimientos en Twitter

Antedecent	Consequent	Supp	Conf
{trust}	*{hillary-clinton}*	0.939688	1
{anger}	*{hillary-clinton}*	0.492217	1
{anticipation}	*{hillary-clinton}*	0.486381	1
{fear}	*{hillary-clinton}*	0.299610	1
{surprise}	*{hillary-clinton}*	0.200389	1
{joy}	*{hillary-clinton}*	0.145914	1
{sadness}	*{hillary-clinton}*	0.079766	1
{disgust}	*{hillary-clinton}*	0.077821	1

Antedecent	Consequent	Supp	Conf
{trust}	*{donald-trump}*	0.945927	1
{anticipation}	*{donald-trump}*	0.594113	1
{surprise}	*{donald-trump}*	0.425051	1
{anger}	*{donald-trump}*	0.345656	1
{fear}	*{donald-trump}*	0.295003	1
{joy}	*{donald-trump}*	0.226557	1
{disgust}	*{donald-trump}*	0.112936	1
{sadness}	*{donald-trump}*	0.074606	1

1. Reglas de asociación

Aplicaciones

IDEAL

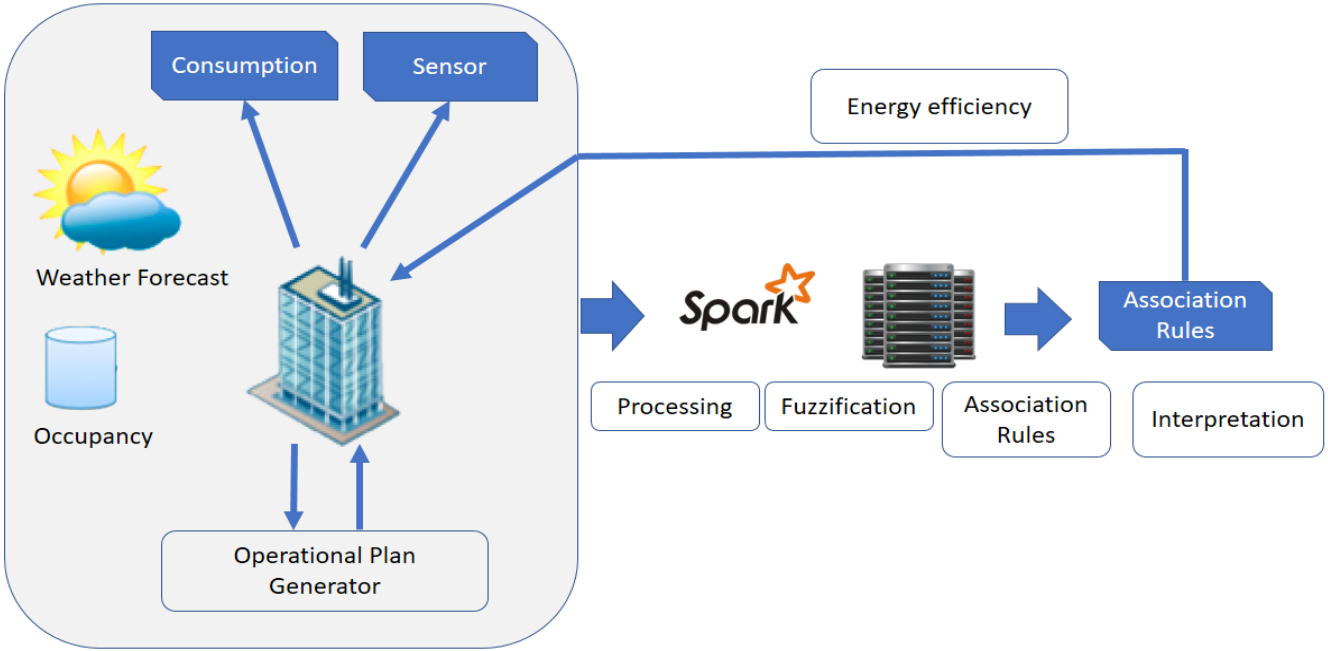
Científicos de la UGR crean un sistema que permite predecir resultados electorales con análisis de opiniones en Twitter

El método ideado en la UGR ofrece de una manera fácilmente interpretable y explicable una serie de relaciones entre conceptos y discusiones en la red social sobre ambos políticos, así como los sentimientos y emociones asociados a los mismos

1. Reglas de asociación

Aplicaciones

Búsqueda de patrones para la mejora de la eficiencia energética



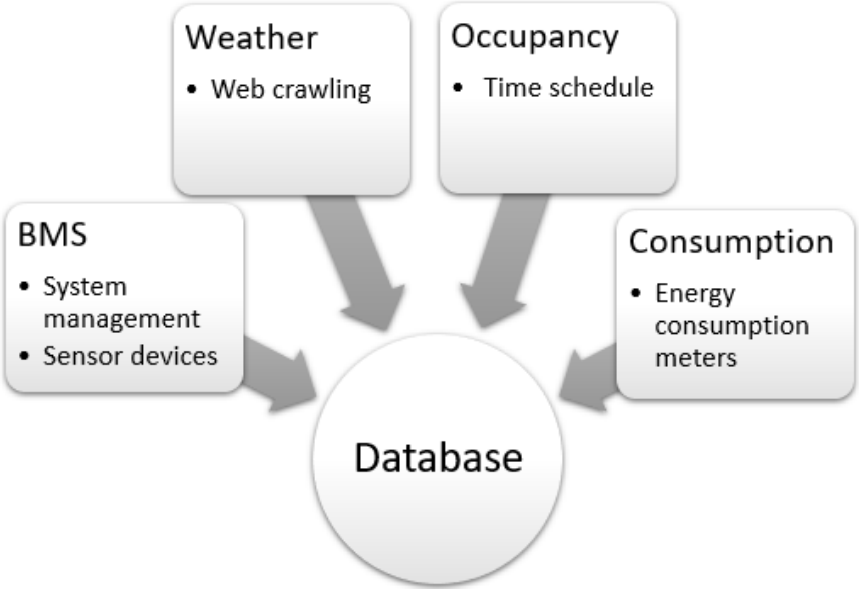
Fernandez-Basso et al. A fuzzy mining approach for energy efficiency in a Big Data framework IEEE Transactions on Fuzzy Systems, 2020.

1. Reglas de asociación

Aplicaciones

Búsqueda de patrones para la mejora de la eficiencia energética

Edificio de oficinas en Bucarest



1. Reglas de asociación

Aplicaciones

Búsqueda de patrones para la mejora de la eficiencia energética

Ejemplos de reglas de asociación difusas obtenidas:

$$\{Setup\ PAN = on,\ Output\ temperature = cold\} \Rightarrow \{PAN\ temperature = confort,\ PAS\ temperature = cold\}$$

Cuando fuera hace frío y el sistema HVAC* en la zona norte (PAN) está encendida entonces se alcanza el confort y en la zona sur (PAS) hace frío

$$\{WindowsPAN = on,\ Windows\ PAS = on\} \Rightarrow \{Output\ temperature = warm;\ Setup\ PAS = off,\ Setup\ PAN = off,\ Temperature = comfort\}$$

Cuando las ventanas están abiertas en la zona norte y sur, entonces el sistema HVAC* no está funcionando, fuera hace calor y dentro se alcanza el confort

*Heating, Ventilation and Air Conditioning

1. Reglas de asociación

Aplicaciones

Búsqueda de patrones para la mejora de la eficiencia energética

SINC

QUIMIOTERAPIA PROTEÍNAS AVES METAMORFOSIS PATENTES

IBEROAMÉRICA OPINIÓN CIENCIA SALUD INNOVA TIERRA TECNOLOGÍA SOCIEDAD

INNOVACIÓN

Un sistema inteligente reduce un 20% el consumo en climatización de edificios

Investigadores de la Universidad de Granada han desarrollado un sistema de control automático para equipos de aire acondicionado que permite reducir en más de un 20% la energía requerida para climatizar grandes edificios no residenciales, como hoteles, oficinas e instalaciones de aeropuertos. Un *software* asociado a este sistema podría estar disponible en el mercado en un par de años.

SINC 𝕏 27/11/2017 11:04 CEST

DESQBRE CIENCIA PARA TI #AndaluciaCiencia

Desarrollan un sistema inteligente que reduce un 20% el consumo de energía en edificios no residenciales

Fuente: Universidad de Granada

IDEAL

Un nuevo sistema inteligente de la UGR reduce un 20% el consumo de energía en grandes edificios

El trabajo de investigación ha sido desarrollado en colaboración con universidades, empresas y centros de investigación de 8 países

sidad de Granada participantes en el proyecto Energy IN TIME. De izquierda a z Basso, Jesús Campaña Gómez, Juan Gómez Romero, María José Martín Bautista, . UGR

Índice

Sesión 4

- Reglas de asociación
  - Definición
  - Medidas
  - Algoritmos
  - Otros tipos de reglas de asociación
  - Aplicaciones

Sesión 5

- Uso de algoritmos de detección de reglas de asociación en Python
- Ejercicio práctico