

RAZONAMIENTO APROXIMADO

José Á. Olivas

MÁSTER EN INTELIGENCIA ARTIFICIAL

Módulo III. Razonamiento aproximado

viu

**Universidad
Internacional
de Valencia**

Este material es de uso exclusivo para los alumnos de la Universidad Internacional de Valencia. No está permitida la reproducción total o parcial de su contenido ni su tratamiento por cualquier método por aquellas personas que no acrediten su relación con la Universidad Internacional de Valencia, sin autorización expresa de la misma.

Edita

Universidad Internacional de Valencia

Máster en
Inteligencia Artificial

Razonamiento aproximado
Módulo III. Razonamiento aproximado
6 ECTS

José Á. Olivas

Leyendas



Enlace de interés



Ejemplo



Importante



Descarga de archivo



abc Los términos resaltados a lo largo del contenido en color **naranja** se recogen en el apartado **GLOSARIO**.

Índice

CAPÍTULO 1. INTRODUCCIÓN AL RAZONAMIENTO APROXIMADO	7
CAPÍTULO 2. MODELOS CLÁSICOS DE TRATAMIENTO DE LA INCERTIDUMBRE	9
2.1. Teorema de Bayes e inferencia bayesiana	10
2.2. Factores de certeza.....	13
2.3. Teoría de la evidencia.....	17
CAPÍTULO 3. LÓGICA BORROSA O DIFUSA.....	23
3.1. Computación suave.....	24
3.2. ¿Qué son los conjuntos borrosos?	25
3.3. La representación borrosa del conocimiento.....	26
3.4. El razonamiento aproximado.....	27
3.5. El éxito del control borroso	29
3.6. Modelo de Mamdani de control borroso	30
3.7. Modelo de Takagi-Sugeno de control borroso	32
3.8. El nuevo reto del razonamiento aproximado: Internet y <i>big data</i>	34
3.8.1. El razonamiento aproximado y la inteligencia artificial	36
3.8.2. Métodos basados en estadística	37
3.8.3. Métodos basados en inteligencia artificial (aprendizaje automático).....	37
3.8.4. Adecuación de los métodos a los problemas	40
CAPÍTULO 4. ALGUNAS APLICACIONES Y EJEMPLOS	42
4.1. Prevención de incendios forestales basada en prototipos deformables borrosos	42
4.1.1. Adquisición de conocimiento y datos	42
4.1.2. Concepto y prototipos (la importancia del científico de datos)	45
4.1.3. Descubrimiento de conocimiento prototípico borroso en datos sobre incendios forestales ...	46
4.2. Computación suave y lógica borrosa para la búsqueda y recuperación de información (en la Web)...	76
4.2.1. Indexación	79
4.2.2. Preprocesado de documentos.....	80

4.2.3. Estructuras de indexación clásicas	82
4.2.4. Consulta	83
4.3. Minería de opiniones.....	95
4.3.1. Principales conceptos	97
4.3.2. Tareas.....	98
4.3.3. Técnicas	99
GLOSARIO.....	103
ENLACES DE INTERÉS	105
BIBLIOGRAFÍA.....	106



Capítulo 1

Introducción al razonamiento aproximado

El **razonamiento aproximado** se enmarca dentro de la **inteligencia computacional** (*computational intelligence* en inglés), esencialmente dentro de lo que se suele denominar **computación suave** (*soft computing* en inglés), por abarcar técnicas tolerantes a la imprecisión y la incertidumbre inherentes a la representación del conocimiento e inferencia de cualquier sistema computacional inteligente. Principalmente la **lógica borrosa o difusa** (*fuzzy logic* en inglés) provee de mecanismos adecuados para formalizar este razonamiento aproximado de una forma completa y precisa.

En este capítulo se introducen los conceptos básicos del razonamiento aproximado, entendido como el conjunto de técnicas que permiten describir, representar y hacer inferencias considerando y aprovechando la imprecisión e incertidumbre inherentes a todos los desarrollos actuales en **inteligencia artificial**. Se presentan inicialmente los modelos clásicos desarrollados para este fin en la inteligencia artificial (teorema de Bayes, factores de certeza y teoría de la evidencia) para posteriormente profundizar en el estudio de la teoría de conjuntos borrosos en sus aspectos principales (representación del conocimiento, razonamiento y control borroso), ilustrándolo con múltiples ejemplos, aplicaciones y herramientas de desarrollo.

Desde que en el año 1965 el profesor **Lotfi A. Zadeh** introdujo la lógica borrosa, muchas han sido sus aplicaciones, las más importantes en el campo del control industrial (lavadoras Bosch con sistema Eco-Fuzzy, ABS de Nissan, aire acondicionado Mitsubishi...). Pero, ya desde sus inicios, la intención

del profesor Zadeh era introducir un formalismo capaz de representar y manipular la incertidumbre e imprecisión inherentes al lenguaje natural y, sobre todo, el **razonamiento aproximado**.

Por otro lado, la mayor parte de la inmensa cantidad de información contenida en Internet está almacenada en documentos textuales en lenguaje natural en multitud de idiomas. Además, muchos de los datos de lo que hoy en día se denomina **big data** tienen una gran carga de imprecisión e incertidumbre.



Capítulo 2

Modelos clásicos de tratamiento de la incertidumbre

El conocimiento humano está lleno de incertidumbre. Los esquemas de representación del conocimiento no contemplan la incertidumbre inherente a la experiencia humana.

Así pues, estos esquemas han de ser complementados con sistemas de representación de la incertidumbre. El conocimiento queda representado por:

- Un esquema de representación.
- Un método de representación de la incertidumbre.



Ejemplo

No es raro que un médico ponga un tratamiento a un paciente a partir de unos síntomas ambiguos. Entendemos frases del lenguaje "entre líneas".

Hay incertidumbre debido a muchas causas:

- Experiencia insuficiente.
- Inadecuada representación del conocimiento.
- Información poco fiable.

- No completitud.
- Inexactitud inherente al lenguaje.

2.1. Teorema de Bayes e inferencia bayesiana



La incertidumbre del conocimiento se puede modelar con probabilidades basándose en la teoría formal de probabilidad del teorema de Bayes. Este permite el cálculo de probabilidades complejas a partir de otras más simples basadas en observaciones reales.

Se ha aplicado en varias áreas de la inteligencia artificial, tales como el reconocimiento de patrones y problemas de clasificación, pero la primera aplicación relevante fue el sistema prospector (Duda, Gaschnig y Hart, 1981).

La probabilidad condicional de los sucesos e y h [$P(h / e)$] se puede interpretar como la relación causa-efecto entre e y h , donde e es la evidencia que soporta la hipótesis h .

Según Bayes, la probabilidad condicional de una hipótesis h , dada la presencia de una evidencia e , viene dada por: $P(\text{hipótesis} / \text{evidencia}) = P(h / e) = P(h \wedge e) / P(e)$ o, de forma equivalente: $P(h / e) = P(h) \times P(e / h) / P(e)$.



Ejemplo

Un médico sabe que la meningitis provoca una rigidez en el cuello de los pacientes, digamos, el 50% de las veces. Conoce también los siguientes hechos incondicionales: la probabilidad *a priori* de que un paciente sufra meningitis es de 1/50,000, y la probabilidad *a priori* de que algún paciente padezca de rigidez del cuello es de 1/20. Si S representa la proposición de que el paciente padezca de rigidez en el cuello y M la proposición de que el paciente tenga meningitis, tenemos que:

$$P(S / M) = 0,5 \quad P(M) = 1 / 50,000 \quad P(S) = 1/20$$

$$P(M / S) = P(S / M) \times P(M) / P(S) = 0,5 \times 1 / 50,000 / 1 / 20 = 0,0002$$



Ejemplo

Se desea conocer si en una cierta localización geográfica habrá un yacimiento de cobre a partir de ciertos estudios de minerales de la zona (evidencias). Se necesita conocer por adelantado las probabilidades de encontrar los minerales y las probabilidades de que haya estos minerales en yacimientos de cobre. Con estos datos podríamos estimar la probabilidad de encontrar un yacimiento de cobre (prospector):

$$P(\text{Cu} / \text{minerales}) \text{ a calcular en función de } P(\text{minerales}) \text{ y } P(\text{minerales} / \text{Cu})$$

La probabilidad condicional de una hipótesis h , dada la presencia de dos evidencias e_1 y e_2 es:

$$P(h/e_1 \wedge e_2) = P(h) \times P(e_1 \wedge e_2/h) / P(e_1 \wedge e_2)$$

La probabilidad condicional de una hipótesis h , dada la presencia de n evidencias $e_1 \dots e_n$ es:

$$P(h/e_1 \wedge \dots \wedge e_n) = P(h) \times P(e_1 \wedge \dots \wedge e_n/h) / P(e_1 \wedge \dots \wedge e_n)$$

En caso de que tengamos k hipótesis mutuamente exclusivas y exhaustivas: $P(h_i \wedge h_j) = 0$ y $\sum_j P(h_j) = 1$. Además, hay n evidencias $e_1 \dots e_n$ independientes $P(e_i/h_j) = P(e_i)$ que las soportan:

$$P(h_i/e_1 \wedge e_2 \dots \wedge e_n) = \frac{P(h_i) \times P(e_1/h_i) \times \dots \times P(e_n/h_i)}{\sum_j P(h_j) \times P(e_1/h_j) \times \dots \times P(e_n/h_j)}$$

j recorre todas las hipótesis

$P(h_i/e_1 \wedge \dots \wedge e_n)$ = probabilidad de h_i dadas las evidencias e_1 a e_n .

$P(h_i)$ = probabilidad de h_i .

$P(e_i/h_i)$ = probabilidad de observar la evidencia e_i dada h_i cierta.



Ejemplo

Sean las evidencias:

e_1 : soltero

e_2 : ingresos altos

e_3 : joven

Estas evidencias soportan las hipótesis:

h_1 : inversor de alto riesgo

h_2 : inversor de bajo riesgo

h_1 y h_2 son mutuamente exclusivas y exhaustivas

$$P(h_1 \wedge h_2) = 0 \text{ y } P(h_1) = 1 - P(h_2)$$

Un experto financiero, viendo sus registros de inversores, puede estimar las probabilidades *a posteriori* siguientes:

$$P(h_1) = 0,3 \quad P(h_2) = 0,7$$

$$P(e_1/h_1) = 0,6 \quad P(e_1/h_2) = 0,4$$

$$P(e_2/h_1) = 0,2 \quad P(e_2/h_2) = 0,8$$

$$P(e_3/h_1) = 0,5 \quad P(e_3/h_2) = 0,5$$

>>>

>>>

Se pretende predecir el perfil de los inversores de mayor y menor riesgo. Las probabilidades *a priori* serán las siguientes:

$$\begin{aligned} P(h_1/e_1) &= [P(h_1)P(e_1/h_1)] / [P(h_1)P(e_1/h_1) + P(h_2)P(e_1/h_2)] = \\ &= 0,3 \times 0,6 / (0,3 \times 0,6 + 0,7 \times 0,4) = 0,39 \end{aligned}$$

$$P(h_1/e_2) = 0,097 \quad P(h_1/e_3) = 0,3$$

$$P(h_2/e_1) = 0,61 \quad P(h_2/e_2) = 0,903 \quad P(h_2/e_3) = 0,7$$

$$P(h_1/e_1 \wedge e_3) = 0,4 \quad P(h_2/e_1 \wedge e_3) = 0,6$$

$$P(h_1/e_1 \wedge e_2 \wedge e_3) = 0,14 \quad P(h_2/e_1 \wedge e_2 \wedge e_3) = 0,86$$

	P_{inicial}	e_1	e_2	e_3	$e_1 \wedge e_3$	$e_1 \wedge e_2 \wedge e_3$
h_1	0,3	0,39	0,097	0,3	0,4	0,14
h_2	0,7	0,61	0,903	0,7	0,6	0,86

h_2 es la situación más probable con la presencia conjunta de $e_1 \wedge e_2 \wedge e_3$.

h_1 es la situación más probable con la presencia conjunta de $e_1 \wedge e_3$.

Defectos del enfoque probabilista

- Se necesita saber un número importante de probabilidades $P(h_i)$ y $P(e_j/h_i)$ y estos valores no son siempre fáciles de conseguir o estimar.
- Las probabilidades requieren muestras representativas.
- Si se descubren nuevas evidencias que condicionan las hipótesis, hay que reconstruir todas las probabilidades.
- La independencia de las hipótesis es difícil que ocurra.
- Este enfoque asume que la presencia de una evidencia a para una hipótesis c también afecta a la negación de esta. Así, si $P(c/a) = 0,8$, entonces $P(\text{no } c/a) = 0,2$, lo cual no es necesariamente cierto en todos los dominios.
- En medicina, la presencia de un síntoma no sería una evidencia que soportase a la vez la existencia y la no existencia de una enfermedad al mismo tiempo.

2.2. Factores de certeza



Los factores de certeza se basan en el **juicio** que tiene un experto sobre el número de ocurrencias de ciertas situaciones o relaciones cuyo conocimiento se desea incluir en una base de conocimientos.

Estas medidas de confianza o factores de certeza **son evaluaciones o apreciaciones personales** de los expertos que añaden al enunciado de su conocimiento. Por ejemplo, si se da A , entonces se dará C casi con toda seguridad.

Se expresan mediante un número o **factor de certeza** (CF). Los factores de certeza no se rigen por probabilidad. No se obtienen de poblaciones muestrales, sino de experiencia. En probabilidad, la suma de la probabilidad de que se dé un hecho y su contrario es 1. Un experto puede sentir que algo es cierto de forma importante, pero puede no saber cuánto de importante es lo contrario.

La teoría de los factores de certeza se introdujo por primera vez en el sistema experto MYCIN (Shortliffe y Buchanan, 1975). La forma típica de usar factores de certeza (MYCIN) es asociándolos a reglas de producción.



Ejemplo

Si la coloración del organismo es grampositiva y la morfología del organismo es coco y la forma de crecimiento del organismo es a base de cadenas, entonces hay una evidencia 0,7 de que la identidad del organismo sea un estreptococo.

0,7 representa el **factor de certeza** asociado a la regla.

El factor de certeza es un valor en el intervalo $[-1, 1]$, donde 1 indica completa confianza y -1 completa no creencia.

Sean dos reglas R_1 y R_2 que alcanzan la misma conclusión h , a partir de dos evidencias e_1 y e_2 diferentes:

R_1 : Si e_1 , entonces h , $CF(h, e_1)$.

R_2 : Si e_2 , entonces h , $CF(h, e_2)$.

El **factor de certeza de h** se calcula como:

- $CF(h, e_1) + CF(h, e_2)(1 - CF(h, e_1))$, si $CF(h, e_1) > 0$ y $CF(h, e_2) > 0$.
- $CF(h, e_1) + CF(h, e_2)(1 + CF(h, e_1))$, si $CF(h, e_1) < 0$ y $CF(h, e_2) < 0$.
- $[CF(h, e_1) + CF(h, e_2)] / (1 - \min(|CF(h, e_1)|, |CF(h, e_2)|))$, en cualquier otro caso.



Ejemplo

R_1 : Si viernes, entonces tráfico 0,8.

R_2 : Si llueve, entonces tráfico 0,7.

Hoy es viernes y llueve. ¿Cuál es la certeza de tráfico? $0,8 + 0,7 \times (1 - 0,8) = 0,94$. Se refuerza la certeza con ambas positivas.

Hoy es viernes y fin de mes. ¿Cuál es la certeza de tráfico? $0,8 + (-0,4) / (1 - \min(0,8; 0,4)) = 0,66$. Se disminuye la fuerza de la mayor.



Ejemplo

Sean h_1 : inversor de alto riesgo

h_2 : inversor de bajo riesgo

e_1 : joven

e_2 : ingresos altos

e_3 : casado

Reglas:

Si joven, entonces inversor de alto riesgo 0,6.

Si ingresos altos, entonces inversor de alto riesgo 0,1

Si casado, entonces inversor de alto riesgo -0,7.

Si se dan e_1 y e_2 , se pueden usar las dos primeras reglas y concluirán ambas h_1 . El valor de certeza de h_1 será:

$$CF(h, e_1 \wedge e_2) = 0,6 + 0,1(1 - 0,6) = 0,64 \text{ (poco aporta la segunda regla)}$$

Si además aparece e_3 , se puede usar la tercera regla, que también concluye h_1 . Entonces su factor de certeza será:

$$\begin{aligned} CF(h, e_1 \wedge e_2 \wedge e_3) &= [CF(h, e_1 \wedge e_2) + CF(h, e_3)] / (1 - \min(|CF(h, e_1 \wedge e_2)|, |CF(h, e_3)|)) = \\ &= [0,64 + (-0,7)] / [1 - 0,64] = -0,16 \end{aligned}$$

Si la persona está casada, se disminuye la certeza del perfil de alto riesgo.

Un factor de certeza se puede asociar no solo a una regla, sino también a una condición de una regla.
Sea:

R_1 : Si e_1 , $CF(e_1)$, entonces h , $CF_R(h, e_1)$.

$CF_R(h, e_1)$ es el factor de certeza de la regla R_1 .

$CF(e_1)$ es el factor de certeza de e_1 .

¿Cuál es la certeza de h con la evidencia e_1 presente?

$$CF(h, e_1) = CF(e_1) \times CF_R(h, e_1)$$

Si una regla tiene varias condiciones enlazadas con una **conjunción** (y) y cada una tiene asociado un factor de certeza:

R_1 : Si e_1 , $CF(e_1)$ y ... y e_n , $CF(e_n)$, entonces h , $CF_R(h, e_1 \wedge \dots \wedge e_n)$

$CF_R(h, e_1 \wedge \dots \wedge e_n)$ es el factor de certeza de la regla R_1

$CF(e_1)$ es el factor de certeza de e_1

.....

$CF(e_n)$ es el factor de certeza de e_n

¿Cuál es la certeza de h con las n evidencias presentes?

$$CF(h, e_1 \wedge \dots \wedge e_n) = CF_R(h, e_1 \wedge \dots \wedge e_n) \times \min[CF(e_1) \dots CF(e_n)]$$

Si una regla tiene varias condiciones enlazadas con una **disyunción** (o) y cada una tiene asociado un factor de certeza:

R_1 : Si e_1 , $CF(e_1)$ o ... o e_n , $CF(e_n)$, entonces h , $CF_R(h, e_1 \vee \dots \vee e_n)$

$CF_R(h, e_1 \vee \dots \vee e_n)$ es el factor de certeza de la regla R_1

$CF(e_1)$ es el factor de certeza de e_1

.....

$CF(e_n)$ es el factor de certeza de e_n

¿Cuál es la certeza de h con las n evidencias presentes?

$$CF(h, e_1 \vee \dots \vee e_n) = CF_R(h, e_1 \vee \dots \vee e_n) \times \max[CF(e_1) \dots CF(e_n)]$$



Ejemplo

R₁: Si A, 0,7 y B, 0,2, entonces C, -0,8.

R₂: Si D, 0,3 o F, 0,6, entonces C, -0,4

Si se dan las evidencias A, B, D y F, ¿cuál es el factor de certeza de C?

$$CF(C, A \wedge B) = 0,2 \times (-0,8) = -0,16$$

$$CF(C, D \vee F) = 0,6 \times (-0,4) = -0,24$$

$$CF(C, (A \wedge B) \wedge (D \vee F)) = -0,16 - 0,24 \times (1 - 0,16) = -0,33$$



Ejemplo

Sean las reglas siguientes:

$$R_1: e_1 \rightarrow h, CF_{R1} = 0,75$$

$$R_2: e_2 \rightarrow h, CF_{R2} = 0,6$$

$$R_3: e_3 \rightarrow h, CF_{R3} = -0,8$$

Donde e_1, e_2, e_3 y h significan lo mismo que en el caso anterior.

Sea otra regla adicional:

$$R_4: h \rightarrow g, CF_{R4} = -0,7$$

g: invertir en bonos

Supongamos que los factores de certeza de e_1, e_2, e_3 son 0,9, 0,5 y 0,9 respectivamente. Supongamos que observamos e_1, e_2, e_3 .

$$CF(h, e_1) = CF(e_1) \times CF_{R1}(h, e_1) = 0,9 \times 0,75 = 0,67$$

$$CF(h, e_2) = CF(e_2) \times CF_{R2}(h, e_2) = 0,5 \times 0,6 = 0,3$$

$$CF(h, e_3) = CF(e_3) \times CF_{R3}(h, e_3) = 0,9 \times (-0,8) = -0,72$$

$$CF(h, e_1 \wedge e_2) = CF(h, e_1) + CF(h, e_2) \times (1 - CF(h, e_1)) = 0,67 + 0,3 \times (1 - 0,67) = 0,77$$

$$CF(h, e_1 \wedge e_2 \wedge e_3) = [CF(h, e_1 \wedge e_2) + CF(h, e_3)] / (1 - \min(|CF(h, e_1 \wedge e_2)|, |CF(h, e_3)|)) = \\ = 0,77 + (-0,72) / (1 - 0,72) = 0,05 / 0,28 = 0,17$$

$$CF(g, h) = CF(h) \times CF_{R4}(g, h) = 0,17 \times (-0,7) = -0,12$$

(No debería invertir)

2.3. Teoría de la evidencia



La **teoría de la evidencia** fue desarrollada por Dempster, y refinada y extendida por Shafer. (De ahí que también se conozca como teoría de Dempster-Shafer). Su objetivo es modelar la incertidumbre del conocimiento eliminando algunos de los puntos flacos del enfoque probabilista o bayesiano. En particular, hace énfasis en que la suma de la creencia en un hecho y en su contrario no tiene por qué ser uno. En probabilidad el aumento de la probabilidad de una hipótesis hace disminuir automáticamente su contraria.

La teoría considera como punto de partida una serie de **entornos** $q_1, q_2 \dots q_n$ que representan el universo. Un entorno cualquiera H consta de un conjunto de **hipótesis** mutuamente exclusivas y exhaustivas que cubren todas las posibles situaciones en el entorno $H = \{h_1, h_2 \dots h_m\}$.

Este conjunto de hipótesis se llama **marco de discernimiento**. Se denomina **conjunto potencia** del marco H al formado por todos los subconjuntos posibles de hipótesis h_i .

(Es algo así como todas las posibles respuestas a las posibles preguntas).

$$P(H) = [\{\emptyset\}, \{h_1\}, \{h_2\} \dots \{h_1, h_2\} \dots \{h_1, h_2 \dots h_m\}]$$

Un marco H de M hipótesis tendrá 2^M subconjuntos representando su conjunto potencia.

La función de **asignación de probabilidad básica o masa de probabilidad** es una función $m: 2^M \rightarrow [0, 1]$ (del conjunto potencia de H al intervalo real $[0, 1]$), definida como:

$$m(\emptyset) = 0$$

$$\sum_{S_i \subseteq H} m(S_i) = 1$$

Donde:

$$0 < m(S_i) < 1$$

S_i es cualquier subconjunto del marco de discernimiento (o, lo que es lo mismo, del conjunto potencia).

$m(S_i)$ es asignada por el experto según su juicio.

La creencia en un conjunto de hipótesis S será:

$$Bel(S) = \sum_{X \subseteq S} m(X)$$

Es la suma de todas las masas de probabilidad de los subconjuntos de S . El valor m asignado a S_i indica la creencia en el conjunto S_i , mientras que la función $Bel(S_i)$ indica la creencia en S_i y todos sus subconjuntos, $Bel(H) = 1$. (La creencia del marco H es 1).

El **intervalo de creencia o certidumbre** de un subconjunto S es: $[Bel(S), 1 - Bel(\sim S)]$, donde $\sim S$ es el conjunto complementario de S con respecto al conjunto potencia del marco de discernimiento.

La **plausibilidad** es una medida del grado en que la evidencia no puede refutar S :

$$\text{Pl}(S) = 1 - \text{Bel}(\sim S)$$

La **incertidumbre** de S o ignorancia es:

$$\text{Pl}(S) = \text{Bel}(S)$$

En el enfoque de Bayes, las probabilidades son un número, mientras que en la teoría de DS las probabilidades se expresan por intervalos de creencia. Cuando los intervalos se estrechan, ambas teorías tienden a coincidir. Segundo Bayes, ¿cuál es la probabilidad de que una persona desconocida hable una lengua extranjera? La respuesta es $\frac{1}{2}$. En el enfoque de DS, esto no tiene por qué ser así, sino que según la evidencia se soportará más o menos la hipótesis de que habla la lengua extranjera o no. Veamos a continuación una comparativa entre ambos enfoques:

Tabla 1

Diferencias entre la probabilidad clásica y la teoría de la evidencia

	Teoría de Probabilidad	Teoría Dempster-Shafer
General	$\sum_i P_i = 1$	$0 < m(\theta) < 1$
Para $S_i \subseteq S$	$P(S_1) \leq P(S_2)$	$m(S_1)$ no es necesariamente $\leq m(S_2)$
Para $S < \bar{S}$	$P(S) + P(\bar{S}) = 1$	Relación entre $m(S)$ y $m(\bar{S})$ no definida



Ejemplo

Un proceso industrial solo puede tener cuatro modos de fallo denominados A, B, C y D . Las masas de probabilidad asignadas por el experto son las siguientes:

$$\begin{aligned}
 m(\emptyset) &= 0 \\
 m(A) &= 0,7 \\
 m(B) &= 0,05 \\
 m(C) &= 0,02 \\
 m(D) &= 0,03 \\
 m(A, B) &= 0,1 \\
 m(A, C) &= 0 \\
 m(A, D) &= 0,05 \\
 m(B, C) &= 0 \\
 m(B, D) &= 0 \\
 m(C, D) &= 0 \\
 m(A, B, C) &= 0 \\
 m(A, B, D) &= 0,05 \\
 m(A, C, D) &= 0 \\
 m(B, C, D) &= 0 \\
 m(A, B, C, D) &= 0
 \end{aligned}$$

>>>

>>>

Creencia de A: $\text{Bel}(A) = m(A) = 0,7$

Creencia de AB: $\text{Bel}(AB) = m(A) + m(B) + m(AB) = 0,85$

Creencia de AD: $\text{Bel}(AD) = m(A) + m(D) + m(AD) = 0,78$

Plausibilidad de A: $\text{Pl}(A) = 1 - m(B) - m(C) - m(D) - m(BC) -$
 $- m(BD) - m(CD) - m(BCD) =$
 $= 1 - 0,05 - 0,02 - 0,03 = 0,9$

Plausibilidad de B: $\text{Pl}(B) = 1 - m(A) - m(C) - m(D) - m(AC) -$
 $- m(AD) - m(CD) - m(ACD) =$
 $= 1 - 0,7 - 0,02 - 0,03 - 0,05 = 0,2$

Plausibilidad de BC: $\text{Pl}(BC) = 1 - m(A) - m(D) - m(AD) =$
 $= 1 - 0,7 - 0,03 - 0,05 = 0,22$

Plausibilidad de ACD: $\text{Pl}(ACD) = 1 - m(B) = 1 - 0,05 = 0,95$

Incertidumbre de A: $\text{Pl}(A) - \text{Bel}(A) = 0,9 - 0,7 = 0,2$

Incertidumbre de B: $\text{Pl}(B) - \text{Bel}(B) = 0,2 - 0,05 = 0,15$

Intervalo de creencia de A: $[0,7; 0,9]$

Intervalo de creencia de ACD: $[0,8; 0,95]$

Conjunto	Creencia	Plausibilidad
A	0,7	0,9
B	0,05	0,2
AB	0,85	0,95
C	0,02	0,02
AC	0,72	0,92
BC	0,07	0,22
ABC	0,87	0,97
D	0,03	0,13
AD	0,78	0,93
BD	0,08	0,28
ABD	0,98	0,98
CD	0,05	0,15
ACD	0,8	0,95
BCD	0,1	0,3
ABCD	1	1

Sea m_1 una función de asignación básica de probabilidad sobre un marco de discernimiento H , y m_2 otra función de asignación básica de probabilidad también sobre H pero con evidencia añadida o actualizada. El resultado de combinar ambas funciones m_1 y m_2 es:

$$m_3(C) = \frac{\sum_{X \cap Y = C} m_1(X) \times m_2(Y)}{1 - \sum_{X \cap Y = \emptyset} m_1(X) \times m_2(Y)}$$

X , Y y C son subconjuntos del marco de discernimiento. El denominador es un factor de normalización para que m esté en $[0, 1]$. A medida que hay más evidencia común, se da mayor soporte a las posibles hipótesis (numerador). La cantidad de creencia medida por

$$\sum_{X \cap Y = \emptyset} m_1(X) \times m_2(Y)$$

mide el **conflicto** existente entre evidencias que soportan las mismas hipótesis. Si es 1, entonces $m_3(C)$ no se puede definir. Es una función de asignación de probabilidad **contradicitoria**.



Ejemplo

Sean cuatro personas en una habitación cerrada: Bob, Jim, Sally y Karen. De repente se va la luz y menos de un minuto después se recupera. En ese momento los miembros de la sala ven que Karen ha sido asesinada víctima de una puñalada. Watson llega a la escena del crimen y deduce que uno de los siguientes hechos, o una combinación, debe de ser cierto:

Bob es culpable

Jim es culpable

Sally es culpable

Watson aplica la teoría de DS y tiene el siguiente marco de discernimiento:

$$H = \{B, J, S\}$$

Su conjunto potencia será:

$$H = [\{\emptyset\}, \{B\}, \{J\}, \{S\}, \{B, J\}, \{B, S\}, \{J, S\}, \{B, J, S\}]$$

Supongamos que Watson asigna las siguientes masas de probabilidad a las diferentes hipótesis:

$m(\{\emptyset\})$	0
$m(\{B\})$	0,1
$m(\{J\})$	0,2
$m(\{S\})$	0,1
$m(\{B, J\})$	0,1
$m(\{B, S\})$	0,1
$m(\{J, S\})$	0,3
$m(\{B, J, S\})$	0,1

>>>

>>>

Creencias:

	<i>m</i>	Bel
{Ø}	0	0
{B}	0,1	0,1
{J}	0,2	0,2
{S}	0,1	0,1
{B,J}	0,1	0,4
{B,S}	0,1	0,3
{J, S}	0,3	0,6
{B,J,S}	0,1	1,0

Watson encuentra unas colillas en un cenicero y unas facturas que antes no había visto y revisa el caso. Esto es, reasigna las masas de probabilidad ante las nuevas evidencias:

$mr(\{\emptyset\})$	0
$mr(\{B\})$	0,2
$mr(\{J\})$	0,1
$mr(\{S\})$	0,05
$mr(\{B,J\})$	0,05
$mr(\{B,S\})$	0,3
$mr(\{J, S\})$	0,1
$mr(\{B,J,S\})$	0,2

¿Cuál sería la nueva masa de probabilidad combinada de ambas?

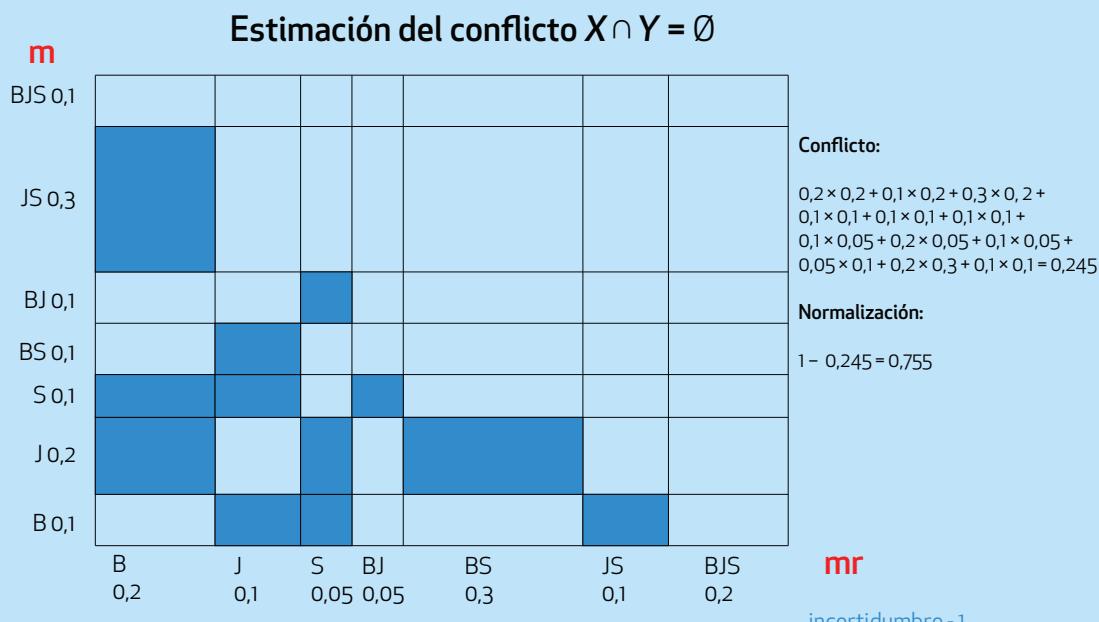
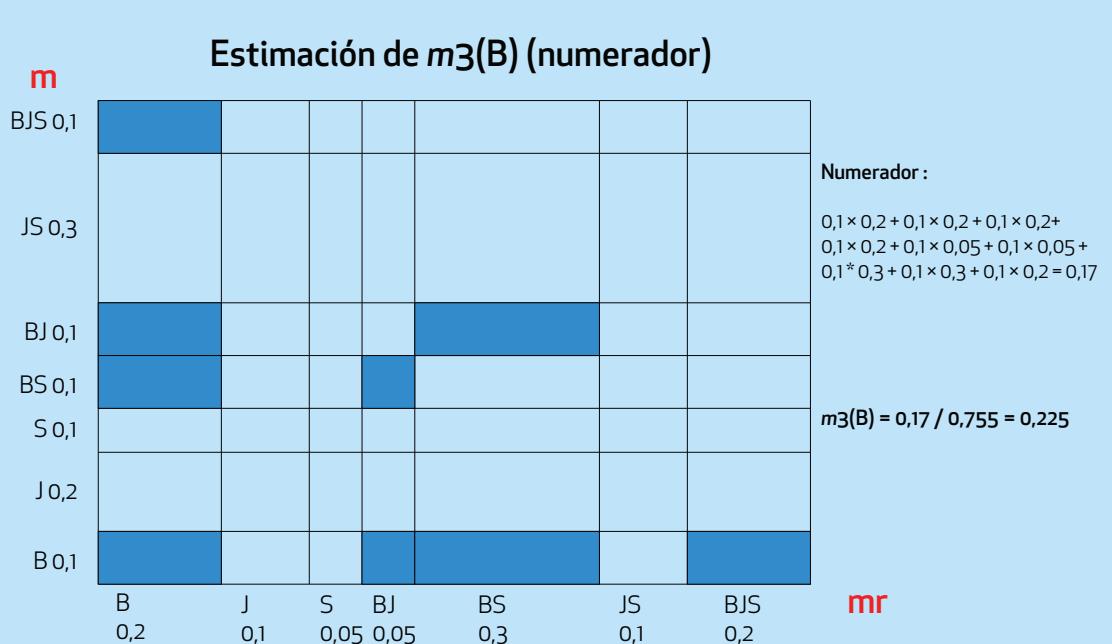


Figura 1. Estimación del conflicto.

>>>

>>>



incertidumbre - 2

Figura 2. Estimación de $m_3(B)$, numerador.

Nuevas masas y creencias combinadas:

	<i>m</i>	Bel
{Ø}	0	0
{B}	0,225	0,225
{J}	0,219	0,219
{S}	0,25	0,25
{B,J}	0,105	0,58
{B,S}	0,04	0,484
{J, S}	0,131	0,6
{B,J,S}	0,03	1,0



Capítulo 3

Lógica borrosa o difusa

La teoría de conjuntos borrosos fue introducida por Lotfi A. Zadeh a mediados de los años sesenta. Previamente, Max Black, en un artículo titulado "Vagueness: An exercise in Logical Analysis" (1937), y Karl Menger, en el artículo "Statistical Metrics" (1942) y los de los años cincuenta sobre relaciones borrosas de indistinguibilidad, sentaron las bases de lo que hoy es una teoría muy utilizada y con buenos resultados.



Bajo el concepto de conjunto borroso (*fuzzy set* en inglés) reside la idea de que los elementos clave en el pensamiento humano no son números sino etiquetas lingüísticas. Estas etiquetas permiten que los objetos pasen de pertenecer de una clase a otra de forma suave y flexible.

La lógica borrosa se puede inscribir en el contexto de la lógica **multivaluada**. En 1922 Lukasiewicz cuestionaba la lógica clásica bivaluada (valores cierto y falso). Además, adelantaba una lógica de valores ciertos en el intervalo unidad como generalización de su lógica **trivaluada**. En los años treinta fueron propuestas lógicas multivaluadas para un número cualquiera de valores ciertos (igual o mayor que 2), identificados mediante números racionales en el intervalo $[0, 1]$.

Uno de los objetivos de la lógica borrosa es proporcionar las bases del razonamiento aproximado, que utiliza premisas imprecisas como instrumento para formular el conocimiento.

3.1. Computación suave

El profesor Zadeh acuñó el término de *soft-computing*, que se puede traducir al español como **computación suave o blanda**.



La computación suave se diferencia de la computación convencional (dura) en que, a diferencia de ella, es tolerante a la imprecisión, la incertidumbre y la verdad parcial. El modelo a seguir para la computación suave es la mente humana y su principio rector es aprovechar la tolerancia a la imprecisión, la incertidumbre y la verdad parcial para lograr la trazabilidad, la robustez y el bajo coste de las soluciones.

Las ideas básicas que subyacen a la computación suave en su estado actual tienen vínculos con muchas **influencias anteriores**, entre ellas los conjuntos difusos introducidos en los años sesenta, los trabajos de los setenta sobre el análisis de sistemas complejos y procesos de decisión, y los de los años ochenta sobre teoría de posibilidades y análisis de datos blandos. La inclusión de la teoría de redes neuronales en la computación suave llegó en un momento posterior. En esta coyuntura, los principales componentes de la computación suave (CS) son la lógica borrosa (LB), la teoría de redes neuronales (RN) y el razonamiento probabilístico (RP), con este último subsumiendo las redes de creencias, los algoritmos genéticos, la teoría del caos y partes de la teoría del aprendizaje.

Lo que es importante tener en cuenta es que el CS no es una mezcla de LB, RN y RP. Se trata más bien de una asociación en la que cada uno de los socios aporta una metodología distinta para abordar los problemas en su ámbito. En esta perspectiva, las principales contribuciones de LB, RN y RP son complementarias y no competitivas.

Implicaciones del soft computing

La complementariedad de LB, RN y RP tiene una consecuencia importante: en muchos casos un problema puede resolverse de forma más eficaz utilizando LB, RN y RP en combinación y no exclusivamente. Un ejemplo llamativo de una combinación particularmente efectiva es lo que se ha llegado a conocer como sistemas neuroborrosos (*neurofuzzy* en inglés).

Estos sistemas son cada vez más visibles como productos de consumo, desde aires acondicionados y lavadoras hasta fotocopiadoras y videocámaras. Menos visibles, pero quizás aún más importantes, son los sistemas neuroprotectores en aplicaciones industriales. Lo que es particularmente significativo es que, tanto en los productos de consumo como en los sistemas industriales, el empleo de técnicas de *soft computing* conduce a sistemas que tienen un alto cociente de inteligencia de máquina (IM).

En gran medida, es el alta IM de los sistemas basados en CS lo que explica el rápido crecimiento en el número y la variedad de aplicaciones de la computación suave y especialmente de la lógica difusa. La estructura conceptual de la computación suave sugiere que los estudiantes deben ser entrenados no solo en teoría de redes neuronales, lógica borrosa o razonamiento probabilístico, sino en todas las metodologías asociadas, aunque no necesariamente en el mismo grado.

Lo mismo se aplica a las revistas, los libros y las conferencias. Hay muchas revistas científicas y libros con este término en su título. Una tendencia similar es perceptible en los títulos de las conferencias y congresos.

3.2. ¿Qué son los conjuntos borrosos?

En un conjunto clásico (*crisp* en inglés), se asigna el valor 0 o 1 a cada elemento para indicar la pertenencia o no a dicho conjunto. Esta función puede generalizarse de forma que los valores asignados a los elementos del conjunto caigan en un rango particular y con ello indiquen el grado de pertenencia de los elementos al conjunto en cuestión. Esta función se llama **función de pertenencia** y el conjunto por ella definida es el **conjunto borroso**. La función de pertenencia μ_A por la que un conjunto borroso A se define, siendo $[0, 1]$ el intervalo de números reales que incluye los extremos, tiene la siguiente forma:

$$\mu_A = X \rightarrow [0, 1]$$

Es decir, mientras que en un conjunto clásico los elementos pertenecen o no pertenecen a él totalmente (por ejemplo, un número puede pertenecer o no al conjunto de los pares, pero no pertenece con un determinado grado), en los conjuntos borrosos hay grados de pertenencia en referencia a un universo local. Por ejemplo, en el contexto de nuestra sociedad actual, una persona de 45 años pertenece al conjunto borroso "viejo" con un grado supongamos de 0,5. Si en vez de usar de referencia nuestra sociedad actual aludíramos a una sociedad donde la esperanza de vida fuera de 40 años, este grado cambiaría.

A es un **subconjunto borroso** de B cuando: $\mu_A(x) \leq \mu_B(x), \forall x \in X$

Originalmente, la teoría de conjuntos borrosos se formuló a partir de un conjunto de operadores también válidos para conjuntos clásicos:

- Negación: $\mu_{\neg A}(x) = 1 - \mu_A(x)$
- Unión: $\mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)]$
- Intersección: $\mu_{A \cap B}(x) = \min[\mu_A(x), \mu_B(x)]$

Posteriormente, se han definido clases de funciones con propiedades axiomáticas adecuadas a la utilidad de cada operador, principalmente las T-normas y T-conormas, que sirven como modelos de la intersección y la unión respectivamente. El origen del uso de las T-normas y T-conormas se remonta a las consecuencias del artículo "Statistical Metrics", de Menger. Para establecer la **desigualdad triangular** (en un triángulo cualquiera, la suma de dos lados siempre es mayor que el tercero), discípulos de Menger establecieron y estudiaron el concepto de norma triangular (T-norma) como operación tipo para componer (sumar probabilísticamente) los lados de un triángulo que no midan un número, sino una función de distribución de probabilidad. Posteriormente, esto se ha revelado como una herramienta adecuada para formalizar la intersección borrosa.

Pero para completar un tipo de razonamiento análogo al que se realiza con lógica clásica, es necesario definir el concepto de implicación. Una implicación borrosa I es en general una función de la forma I: $[0, 1] \times [0, 1] \rightarrow [0, 1]$.

Para cualesquiera dos valores ciertos a y b de proposiciones borrosas, p y q definen el valor cierto $I(a, b)$ de la proposición condicional "si p , entonces q ". Es una extensión de la implicación clásica $p \rightarrow q$ del dominio restringido $\{0, 1\}$ al dominio completo $[0, 1]$.

Así como en lógica clásica una implicación se puede expresar de distintas formas y todas son equivalentes, sus extensiones a lógica borrosa resultan no ser equivalentes y han dado lugar a diferentes clases de implicaciones borrosas.

Por último, existe un principio que permite la generalización de conceptos matemáticos nítidos (*crisp* en inglés) a la teoría de conjuntos borrosos. Cualquier función que asocie puntos $x_1, x_2 \dots x_n$ del conjunto *crisp* X al Y puede generalizarse de forma que asocie subconjuntos borrosos de X en Y . Es el denominado **principio de extensión**.

3.3. La representación borrosa del conocimiento

En lenguaje natural se describen objetos o situaciones en términos imprecisos: grande, joven, tímido... El razonamiento basado en estos términos no puede ser exacto, ya que normalmente representan impresiones subjetivas, quizás probables, pero no exactas. Por ello, la teoría de conjuntos borrosos se presenta más adecuada que la lógica clásica para representar el conocimiento humano, ya que permite que los fenómenos y observaciones tengan más de dos estados lógicos.

Para usar la construcción de conjuntos borrosos en sistemas inteligentes, son necesarias técnicas específicas de adquisición de conocimiento. Las más usadas son las entrevistas y los formularios, pero parece adecuado adaptar otras técnicas al campo borroso.

En los **sistemas basados en el conocimiento** la función de pertenencia debe ser obtenida del experto en ese dominio de conocimiento. Esta función no ha de ser confundida con una función de distribución de probabilidad basada en la repetición de las observaciones, sino en la opinión del experto.

La representación habitual del conocimiento en términos borrosos se realiza por medio de reglas, del tipo:

$$\begin{aligned} \text{Si } & x_1 \text{ es } A_{1,1} \\ & \text{y/o } x_2 \text{ es } A_{2,1} \\ & \text{y/o } x_n \text{ es } A_{i,n} \\ \text{Entonces, } & y \text{ es } B_i \end{aligned}$$

Cada variable que interviene como hipótesis en una regla tiene asociado un dominio. Cada dominio puede estar dividido en tantos conjuntos borrosos como el experto considere oportuno. Cada una de estas particiones tiene asociada una **etiqueta lingüística**.

Un **conjunto de términos** (*term set* en inglés, véase la Figura 3) es un conjunto finito, prioritariamente con 7 ± 2 elementos, que son restricciones de una variable lingüística borrosa. Este conjunto de elementos debe ser suficiente para describir cualquier situación relativa al contexto en el que se sitúa el problema.

Por ejemplo, el siguiente conjunto de términos pretende reflejar una descripción estándar de lo que se entiende por altura (referida a personas):

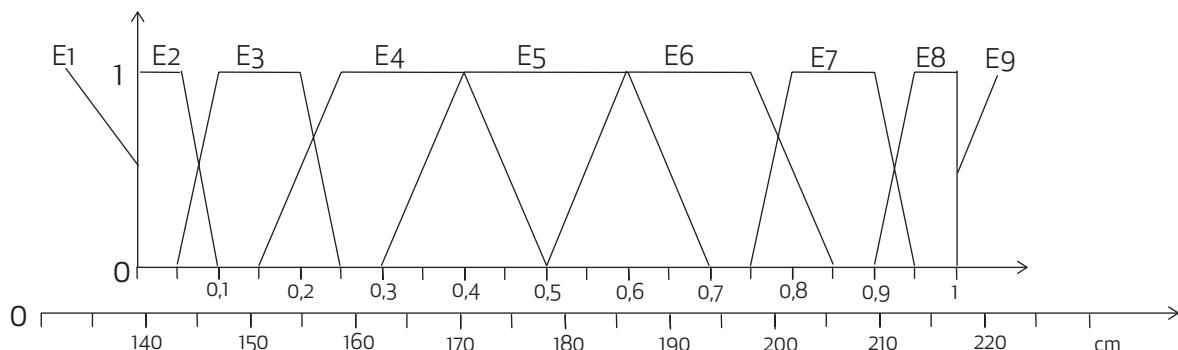


Figura 3. Ejemplo de conjunto de términos.

Valores lingüísticos	(a, b, c, d)
E1: Bajísimo	(0, 0, 0, 0)
E2: Muy bajo	(0, 0, 0, 05, 0, 1)
E3: Bastante bajo	(0, 05, 0, 1, 0, 2, 0, 25)
E4: Ligeramente bajo	(0, 15, 0, 25, 0, 4, 0, 5)
E5: Normal	(0, 3, 0, 4, 0, 6, 0, 7)
E6: Ligeramente alto	(0, 5, 0, 6, 0, 75, 0, 85)
E7: Bastante alto	(0, 75, 0, 8, 0, 9, 0, 95)
E8: Muy alto	(0, 9, 0, 95, 1, 1)
E9: Altísimo	(1, 1, 1, 1)

El universo de discurso (alturas) está normalizado entre 0 y 1 aunque refleja, por ejemplo, entre 130 y 230 cm.

3.4. El razonamiento aproximado



Zadeh introdujo la teoría del razonamiento aproximado y otros muchos autores han hecho contribuciones importantes a este campo. Aunque superficialmente pueda parecer que la teoría del razonamiento aproximado y la lógica clásica se diferencian enormemente, la lógica clásica puede ser vista como un caso especial de la primera. En ambos sistemas, se pueden ver las premisas como inductoras de subconjuntos de mundos posibles que las satisfacen, aunque en el caso de la teoría del razonamiento aproximado esos conjuntos son subconjuntos borrosos. La inferencia entre ambos sistemas está basada en una regla de inclusión: una hipótesis se infiere de una colección de premisas si el subconjunto de mundos posibles que satisfacen la conjunción de las premisas está contenido en el subconjunto de mundos posibles que satisfacen la hipótesis.

>>>

>>>

La contribución fundamental del razonamiento aproximado es el uso que hace de las variables y la representación de las proposiciones en términos de valores de verdad lingüísticos (subconjuntos borrosos) como valores de esas variables. La lógica clásica solo usa de modo implícito la idea de variable, en el sentido de valor de verdad asociado a una proposición. Sin embargo, su naturaleza binaria le permite ocultar este hecho, ya que nos podemos referir a una proposición que es verdadera por su denotación, p , y a una que es falsa simplemente por su negación, $\neg p$, evitando así la introducción de una variable Vp cuyo valor sea la valoración de la proposición p . El uso del concepto de variable en la teoría del razonamiento aproximado conduce a tratar dominios que no están dentro del ámbito de la lógica clásica, como es el caso de los problemas que tratan los sistemas expertos borrosos o los controladores borrosos.

La teoría del razonamiento aproximado permite representar también cuantificadores lingüísticos situados entre el “para todo” y el “existe” clásicos. Esto facilita representar enunciados como “La mayoría de los coches lujosos son caros” o “Bastantes electores votarán en blanco”. Zadeh indicó que un cuantificador como “la mayoría” puede ser representado como un subconjunto borroso sobre un universo de discurso. Los cuantificadores aproximados se usan para representar conocimiento de sentido común.

Una extensión interesante de la teoría del razonamiento aproximado es la posibilidad de tratar con ella conocimiento prototípico. Reiter sugirió una aproximación a la representación de conocimiento de sentido común usando reglas por defecto, y Yager lo estudió en el marco de la teoría del razonamiento aproximado. De acuerdo con Reiter, una regla por defecto, tal como “típicamente los pájaros vuelan”, puede ser interpretada así: si un objeto es un pájaro y nuestro conocimiento disponible no es incompatible con que el objeto vuela, entonces asumimos que el pájaro vuela.

La lógica binaria puede ser vista como un caso especial de la teoría del razonamiento aproximado en el cual los conjuntos base tienen dos elementos $\{T, F\}$ y los grados de pertenencia se restringen a 1 a 0. La lógica posiblística puede ser vista como una extensión de esta, ya que, aunque se restringen los conjuntos base de valores a dos, T y F , se permite que los grados de pertenencia sean números en el intervalo unidad.

La lógica borrosa extiende la lógica binaria, lo que permite su formalización en términos de la teoría del razonamiento aproximado. Así, “ p es verdadero” alcanzaría la representación “ Vp es $\{1/T, 0/F\}$ ”; “ p es falso” alcanzaría la representación “ Vp es $\{0/T, 1/F\}$ ”; y “ Vp es $\{1/T, 1/F\}$ ” indica que el valor de verdad de la proposición es desconocido. En cualquiera de los casos, el conjunto base asociado a la variable valor de verdad de la proposición p es $\{T, F\}$.

La regla principal de inferencia en lógica clásica, modo de razonamiento ya introducido por los megáricos y estoicos en tiempos de Aristóteles, es el *modus ponens* (nombre asignado en la edad media), que consiste en que, si se tiene la regla $A \rightarrow B$ y se da el hecho A , se puede concluir B . Por ejemplo, si la regla es “Si llueve, entonces me mojo” y se da el hecho cierto de que “llueve”, entonces podré concluir que “me mojo”.

En lógica borrosa se puede generalizar esta regla, quedando su esquema de la siguiente forma:

Regla: Si x es A , entonces y es B .

Hecho: x es A'

Conclusión: y es B'



Ejemplo

Por ejemplo, la regla podría ser "Si la ciudad es grande (x es A), el tráfico es muy denso (y es B)"; el hecho podría ser "La ciudad no es muy grande (x es A')". ¿Qué se podría decir del tráfico ($B'(x)$)?

Supongamos que las variables están relacionadas no necesariamente por una función, sino por cualquier relación. Supongamos que es una relación binaria borrosa R en el universo $X \times Y$. A' y B' son conjuntos borrosos en X e Y respectivamente. Si conocemos R y A' , podríamos conocer B' mediante la **regla compositinal de inferencia**:

$$B' = A'(x)^\circ R(x, y)$$

$$B'(y) = \sup_{x \in X} \min[A'(x), R(x, y)]$$

Donde $R(x, y) = I(A(x), B(y))$ (función de implicación)

3.5. El éxito del control borroso

Aunque la intención original del profesor Zadeh era crear un formalismo para manipular de forma más eficiente la imprecisión y la vaguedad del razonamiento humano expresado lingüísticamente, causó cierta sorpresa que el éxito de la lógica borrosa llegase en el campo del control automático de procesos. Esto se debió principalmente al auge de lo borroso en Japón, iniciado en 1987 y que alcanzó su máximo apogeo a principios de los noventa. Desde entonces, han sido infinidad los productos lanzados al mercado que usan tecnología borrosa, muchos de ellos utilizando la etiqueta *fuzzy* como símbolo de calidad y prestaciones avanzadas (podemos ver en televisión el anuncio publicitario de la lavadora Bosch con sistema Eco-Fuzzy).

En 1974, el profesor Mamdani experimentó con éxito un controlador borroso en una máquina de vapor, pero la primera implantación real de un controlador de este tipo fue realizada en 1980 por F. L. Smidt & Co. en una planta cementera en Dinamarca. En 1983, Fuji aplicó la lógica borrosa para el control de inyección química para plantas depuradoras de agua, por primera vez en Japón. En 1987 la empresa OMRON desarrolló los primeros controladores borrosos comerciales con el profesor Yamakawa.

A partir de ese momento, el control borroso ha sido aplicado con éxito en muy diversas ramas tecnológicas: la metalurgia, robots de fabricación, controles de maniobra de aviones, ascensores y trenes (tren-metro de Sendai, Japón, 1987), sensores, imagen y sonido (sistema de estabilización de imagen en cámaras fotográficas y de vídeo Sony, Sanyo, Canon...), electrodomésticos (lavadoras de Panasonic o Bosch, aire acondicionado Mitsubishi, rice-cooker...), automoción (sistemas de ABS de Mazda y Nissan, cambio automático de Renault, control automático de velocidad, climatizadores...) y una larga lista de aplicaciones comerciales.

Pero ¿dónde radica el éxito de las aplicaciones de control? El éxito radica en la simplicidad, tanto conceptual como de desarrollo. Los dos paradigmas clásicos de control borroso son el enfoque de Mamdani y el de Takagi-Sugeno que se describen brevemente a continuación.

3.6. Modelo de Mamdani de control borroso

En el enfoque de **Mamdani** un experto ha de especificar su conocimiento en forma de reglas lingüísticas, debe definir las etiquetas lingüísticas que van a describir los estados de las variables. Para cada entrada ($X_1, X_2 \dots X_n$) se ha de especificar la correspondiente etiqueta lingüística que define la salida Y . Cada una de las n variables de entrada y la de salida han de repartirse en conjuntos borrosos (*term sets*) específicos con unos significados, similares a los que se han presentado en este manual. Así podrán ser definidos P_i , conjuntos borrosos distintos en la variable X_i . Lo mismo se puede hacer con el resto de las variables y la salida. Cada conjunto borroso P_i ha de llevar asociado una etiqueta lingüística.

En la base de conocimiento las reglas tienen la forma clásica:

$$\text{Si } h_1 \text{ es } A^{(1)} \text{ y } h_2 \text{ es } A^{(2)} \dots \text{ y } h_n \text{ es } A^{(n)}, \text{ entonces } \eta \text{ es } B.$$

$A^{(1)} \dots A^{(n)}$ y B son etiquetas lingüísticas que corresponden a los conjuntos borrosos $\mu_{(1)} \dots \mu_{(n)}$ y μ , de acuerdo a las particiones de los conjuntos $X_1, X_2 \times \dots \times X_n$ e Y .

La base de reglas constará de K reglas de control.

La lógica de control consiste en comprobar separadamente cada regla de la base de reglas. Se determina el grado de cumplimiento de cada hipótesis de la regla de acuerdo a la variable medida. Si h_1 es $A^{(1)}$ y ... y h_n es $A^{(n)}$, entonces η es B . Para cada regla, se observa el grado de compatibilidad de las variables medidas realmente $x_1, x_2 \dots x_n$ con las etiquetas lingüísticas $A^{(1)} \dots A^{(n)}$ y después se hace la conjunción de grados de cumplimiento.

Para cada regla R_r de las K de control se calcula: $\alpha_r = \min\{\mu_{(1r)} \dots \mu_{(nr)}\}$. La salida de R_r es un conjunto borroso de valores de salida obtenidos cortando el conjunto borroso μ_{ir} asociado con la conclusión de la regla R_r en el nivel de cumplimiento α_r .

Supongamos, por ejemplo, una base de reglas como la siguiente:

R_1 : Si el ángulo A es positivo pequeño y el ángulo B es aproximadamente 0, entonces el ángulo de salida es positivo pequeño.

R_2 : Si el ángulo A es positivo medio y el ángulo B es aproximadamente 0, entonces el ángulo de salida es positivo medio.

Las variables de entrada (ángulo A y ángulo B) y la de salida (ángulo de salida) tienen cada una asignada un *term set*. Supongamos que los datos reales medidos son los siguientes: ángulo A = 36° y ángulo B = $-2,25^\circ$. ¿Cuál debe ser la salida (orden) que debe dar el controlador borroso? La evaluación de la regla R₁ es la siguiente:

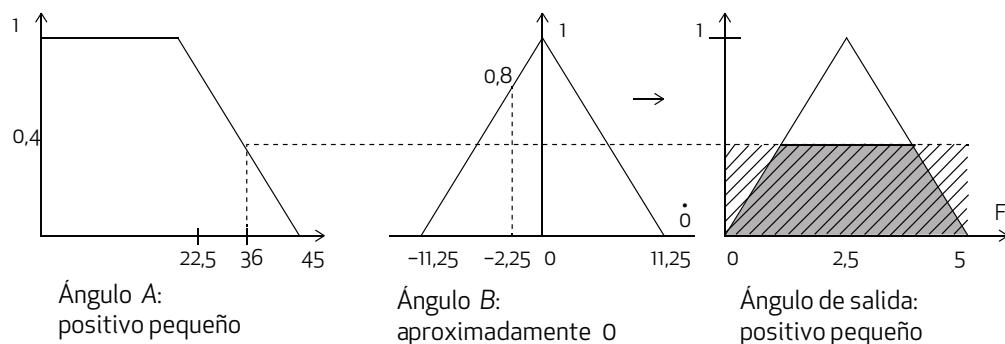


Figura 4. Evaluación de R₁.

Y la evaluación de la regla R₂ es:

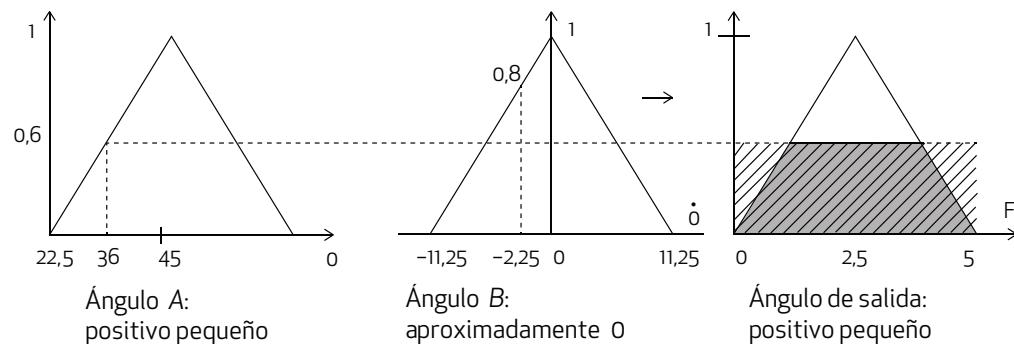


Figura 5. Evaluación de R₂.

Tras la evaluación de cada regla, se han de combinar todos los conjuntos borrosos obtenidos de la salida de las reglas mediante la operación máximo (unión):

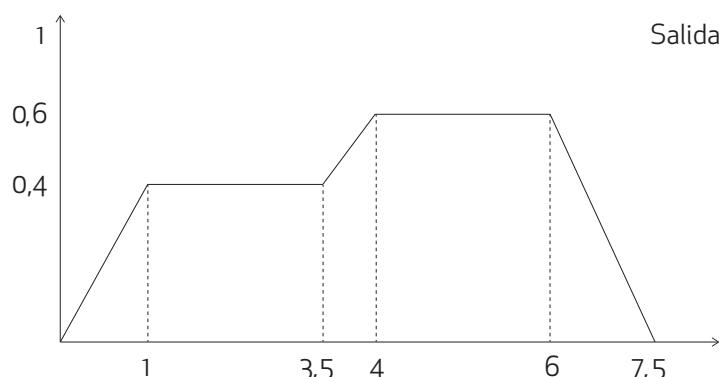


Figura 6. Salida de las reglas.

La salida es la asociación de cada tupla de entradas medidas $(x_1 \dots x_n) \in X_1 \times \dots \times X_n$ con un conjunto borroso de salida para Y . Sin embargo, el sistema a controlar no entendería un conjunto borroso como orden, sino que necesita un valor concreto para actuar, en nuestro ejemplo un ángulo de salida. Por ello, es necesario un interfaz de **defuzzificación o desborrosificación**, que puede seguir varias estrategias: usar algún valor dentro del máximo del conjunto de salida (en el ejemplo, cualquier valor en $[4^\circ, 6^\circ]$ podría ser el valor de salida), usar la media de los máximos (con este criterio, en el ejemplo el valor de salida sería 5°) o calcular la proyección sobre el eje X del centro de gravedad del conjunto borroso de salida (en el ejemplo el valor de salida con este método es $3,9^\circ$). Cada uno de los métodos de desborrosificación presenta sus ventajas e inconvenientes.

3.7. Modelo de Takagi-Sugeno de control borroso

En el enfoque de **Takagi-Sugeno** se mantiene la misma especificación de las particiones borrosas de los dominios de las entradas que en el modelo de Mamdani, pero no se requiere una partición borrosa del dominio de salida. Las reglas de control deben contener una función f_r de $X_1 \times \dots \times X_n$ en Y , que se supone generalmente lineal:

$$f_r(x_1 \dots x_n) = a_1^{(r)} x_1 + \dots + a_n^{(r)} x_n + a^{(r)}$$

$$R_r: \text{si } h_1 \text{ es } A_{i_{1,r}}^{(1)} \text{ y } \dots \text{ y } h_n \text{ es } A_{i_{n,r}}^{(n)}$$

$$\text{Entonces } \eta = f_r(h_1, \dots, h_n)$$

El grado de aplicabilidad a_r se obtiene de la misma manera que el modelo de Mamdani y el valor de control de salida se obtiene como:

$$\eta = \frac{\sum_{r=1}^k \alpha_r f_r(x_1, \dots, x_n)}{\sum_{r=1}^k \alpha_r}$$

Supóngase, por ejemplo, que el proceso de secado de un producto se realiza mediante un ventilador cuya velocidad se regula según la temperatura del producto. El control de la velocidad del ventilador se realiza utilizando un controlador borroso basado en el enfoque de Takagi-Sugeno. El universo de discurso para la variable temperatura es $[0,70] (\text{°C})$. Sobre ese universo de discurso se definen los siguientes conjuntos borrosos:

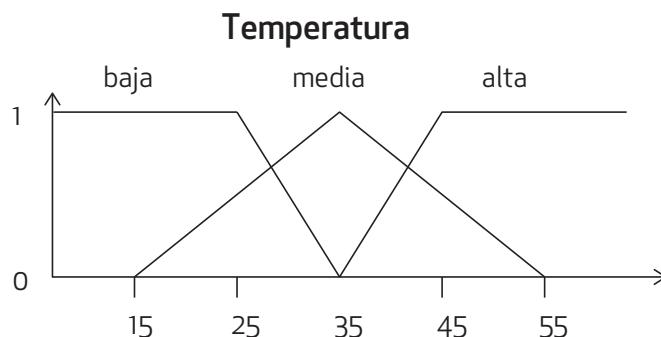


Figura 7. Definición de los conjuntos borrosos.

La base de conocimientos que utiliza el controlador es la siguiente:

REGLA	Temperatura	p_0	p_1
R_1	Alta	700	500
R_2	Media	100	200
R_3	Baja	100	50

"Temperatura" es la hipótesis de las reglas y p_0 y p_1 son los coeficientes de la función consecuente de cada regla que define la velocidad del ventilador. En un caso real se observa un producto con una temperatura de 40 °C. ¿Cuál será la velocidad del ventilador según un controlador borroso basado en el enfoque de Takagi-Sugeno?

Las reglas que se disparan son R_2 y R_3 (R_1 no se dispara porque la pertenencia de 40 °C al conjunto borroso "Temperatura baja" es nula). Para las reglas R_2 y R_3 ocurre lo siguiente:

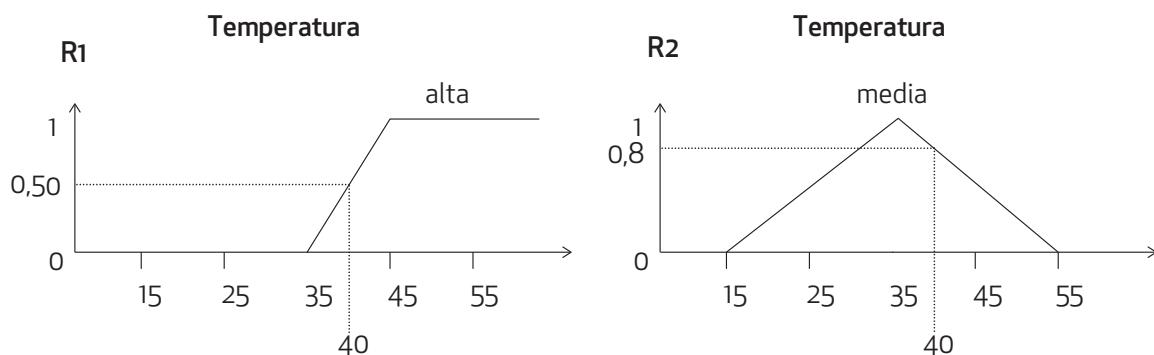


Figura 8. Disparo de reglas R_1 y R_2 .

Así pues, $f_{R1} = 700 + 500 \times 40 = 20.700$ y $f_{R2} = 100 + 200 \times 40 = 8.100$. La velocidad del ventilador, v , resultante será:

$$v = \frac{\alpha_{R1} \times f_{R1} + \alpha_{R2} \times f_{R2}}{\alpha_{R1} + \alpha_{R2}} = \frac{0,5 \times 20.700 + 0,75 \times 8.100}{0,5 + 0,75} = 14.760,6 \text{ rpm}$$

La ventaja más importante de este paradigma es que no es necesaria etapa de desborrosificación. Sin embargo, a veces hay problemas importantes para conseguir los coeficientes de los consecuentes de las reglas en la base de conocimientos.

Pese a las limitaciones e inconvenientes que puedan presentar ambos modelos, lo que sí parece claro es que su simplicidad y buenos resultados son los principales motivos del éxito que ha tenido el control borroso.

3.8. El nuevo reto del razonamiento aproximado: Internet y big data

Los sistemas de ayuda a la decisión (*decision support systems* en inglés) y todas sus múltiples variantes son sistemas computacionales que proporcionan consejos para la mejora en la toma de decisiones. Básicamente, estos consejos pueden venir de tres tipos de análisis:

- **Análisis descriptivo.**

Es un resumen claro y fácil de entender de una colección de datos. Es el fundamento y concepto más básico de todas las estadísticas. Si visualizan los datos para entender el pasado y el presente. Se describen los datos con tablas o gráficos. Las descripciones de la variabilidad y la posición son numéricas.



Figura 9. Diferentes tipos de visualización de datos.

- **Análisis predictivo.**

Se extrapolan funciones (tendencia para el futuro, pero no hay capacidad de pronóstico, hechos/cambios puntuales). Existen correlaciones entre variables (demasiado evidentes, no suelen funcionar de forma muy fina).

Consiste en encontrar **patrones** en los datos que puedan ser aplicados a situaciones futuras. Algunas de las técnicas que se utilizan son el **descubrimiento de conocimiento en bases de datos** (KDD, *knowledge discovery in databases* en inglés) y la **minería de datos** (*data mining*), así como métodos de **agrupamiento** (*clustering*) y **clasificación**.

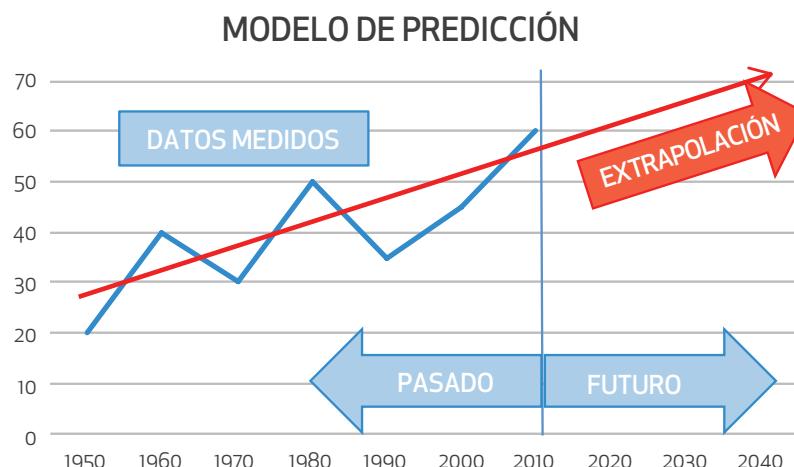


Figura 10. Extrapolación de funciones (por ejemplo, estimaciones o líneas de tendencia).

- **Análisis prescriptivo.**

El análisis predictivo se centraba en un escenario futuro, mientras que el prescriptivo se centra en múltiples alternativas. Por lo tanto, un modelo prescriptivo puede ser considerado como una **combinación de modelos predictivos** (uno por cada posible escenario), que se ejecutan en paralelo. El objetivo es encontrar la mejor opción posible: la **optimización**. Un ejemplo podría ser la elección del tratamiento más adecuado en oncología. Las técnicas usadas son técnicas de investigación operativa, algoritmos genéticos, técnicas estocásticas, metaheurísticas, etc.

Conviene distinguir entre **predicción** y **pronóstico**. La primera tiene que ver con la estimación, con la extrapolación, la continuidad... Por ejemplo, podemos estimar (predecir) cuántos habitantes tendrá Madrid en 2025 partiendo de la tendencia demográfica. En cambio, el pronóstico consiste en anticiparse a un hecho puntual partiendo de un conjunto pequeño de alternativas. Por ejemplo, en una quiniela de fútbol pone "Marque con una X su pronóstico" y las alternativas son 1-X-2, o un terremoto se debe pronosticar partiendo de las alternativas "sí" y "no". Los sistemas para abordar una u otra opción son claramente distintos. Los primeros se basan en la extrapolación y los segundos en una serie de características para anticiparse al hecho puntual (por ejemplo, en las quinielas saber si alguno de los equipos se juega algo importante como el descenso, si hay un jugador importante lesionado o quién va a ser el árbitro del partido).

Hay una gran cantidad de aplicaciones de este tipo de sistemas. Si nos centramos en el **razonamiento aproximado** en sistemas de medicina, podemos citar tres pequeños ejemplos que se pueden estudiar en detalle en las referencias de la bibliografía recomendada, desarrollados en el marco del grupo de investigación liderado por el autor de este tema.



Ejemplo

El primero tiene que ver con el concepto de *enfermedades borrosas* (*fuzzy deseases* en inglés), implantado en un sistema para diagnosticar y tratar fibromialgia, que obtuvo el premio al mejor trabajo en el congreso de la Asociación Española para la Inteligencia Artificial en 2015 (Romero, Olivas, Romero, Alonso y Serrano, 2017). En este trabajo se propone el concepto de **prototipo deformable borroso** para caracterizar enfermedades que pueden ser confundidas o enmascarar otras, incluso que no están perfectamente caracterizadas o asumidas por toda la comunidad médica.



Ejemplo

En el segundo ejemplo se muestra el diseño de un sistema de ayuda a la decisión en oncología a partir de la detección, la clasificación y el uso de las expresiones causales y condicionales en textos médicos (en este caso de Mayo Clinic y Mount Sinai Hospital). En la Figura 11 se puede ver el grafo causal de la pregunta “¿Qué causa cáncer de pulmón?”:

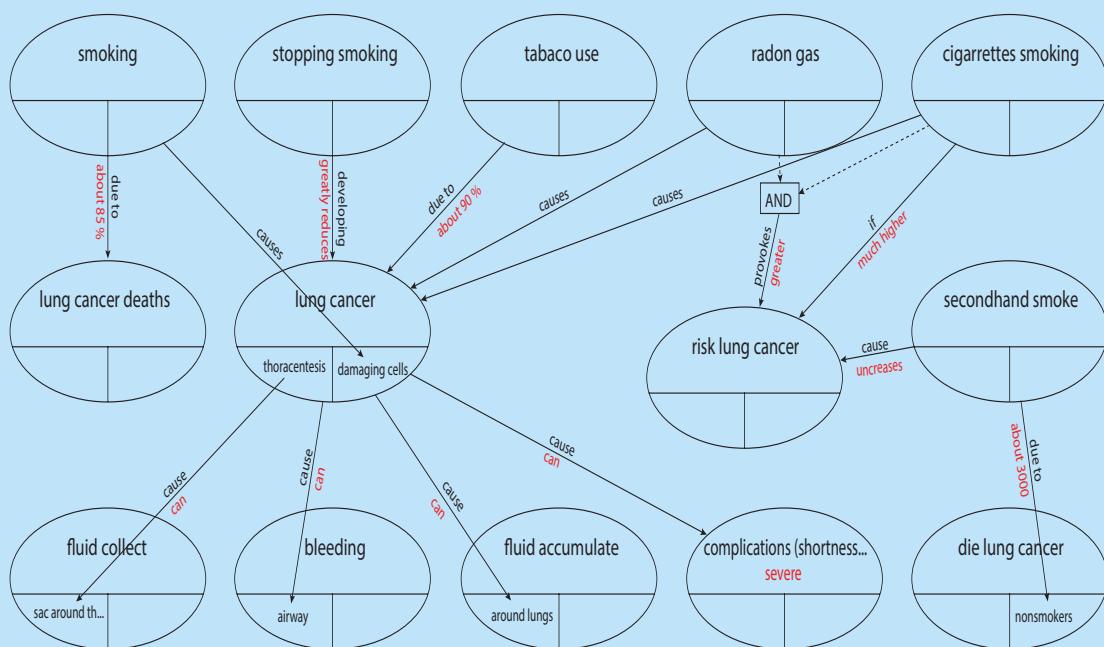


Figura 11. Grafo causal de la pregunta “¿Qué causa cáncer de pulmón?”. Adaptado de “Extracting Answers from causal mechanisms in a medical document”, por A. Sobrino, C. Puente y J. Á. Olivas, 2014, *Neurocomputing*, 135, pp. 53-60.



Ejemplo

En el tercer caso se presenta un estudio inicial sobre el desarrollo de un sistema para la estimación de probabilidades de sufrir determinados tipos de cáncer a partir del estudio del genograma del paciente (Calatrava, Oruezábal, Olivas, Romero y Serrano, 2015).

3.8.1. El razonamiento aproximado y la inteligencia artificial

Hoy en día el **científico de datos** (*data scientist* en inglés) es una figura muy demandada y escasa en ambientes profesionales, científicos o académicos que debe poseer conocimientos de computación, bases de datos, inteligencia artificial, **aprendizaje automático**, estadística, visualización, reconocimiento de patrones, sociología, psicología, KDD y minería de datos. El científico de datos es el encargado de seleccionar y guiar las herramientas y técnicas más adecuadas para cada problema y los objetivos concretos.

Ante la dificultad de encontrar profesionales con este perfil tan completo, la mejor forma de paliar esta dificultad debería ser la formación de equipos multidisciplinares que cubran todas estas necesidades. Sin embargo, esto tampoco es frecuente y por ello gran parte de los proyectos de análisis de datos que llevan a cabo diferentes instituciones y entidades no consiguen los resultados que podrían ser esperados.

Otro problema frecuente e importante a la hora de desarrollar este tipo de proyectos es la tendencia a aplicar ciegamente las herramientas, los métodos o los algoritmos que el equipo mejor conoce ("al que solo dispone de un martillo, todos los problemas le parecen clavos") sobre las bases de datos de las que se dispone (normalmente solo una), sin importar en principio cuál es el propósito final de ese análisis de datos. Esta aproximación errónea se podría mejorar en gran medida disponiendo de un buen científico de datos, en el sentido anteriormente descrito, capaz de guiar todo el proceso con criterios bien sustentados.

3.8.2. Métodos basados en estadística

Las técnicas de **regresión** expresan la formalización de una relación significativa entre dos o más variables para calcular pronósticos a partir del conocimiento de los valores en un individuo concreto. Entre otros tipos de regresión, para el análisis de datos se suelen usar principalmente las siguientes:

- Regresión lineal (aproximación de la dependencia entre una variable dependiente y variables independientes).
- Regresión múltiple (para predecir el valor de una variable dependiente a partir de variables independientes).
- Regresión logística (para predecir variables categóricas).
- CART (*classification and regression trees*, introducidos por Leo Breiman).

Además, también se suelen usar otras técnicas clásicas en el ámbito de la estadística, entre otras:

- Técnicas de **extrapolación** de funciones.
- Técnicas de **aproximación y ajuste** de funciones.
- Técnicas de **agrupamiento** basadas en medidas estadísticas (*clustering*).

Es importante reseñar que muchas se pueden englobar tanto en técnicas estadísticas como de aprendizaje automático, es decir, la mayoría de las técnicas de aprendizaje automático se basan en mecanismos estadísticos.

3.8.3. Métodos basados en inteligencia artificial (aprendizaje automático)

La **inteligencia artificial** (IA) se puede ver como la disciplina del ámbito de la computación y los sistemas de información que pretende simular computacionalmente comportamientos humanos que pueden ser considerados como inteligentes. Hay diversas ramas dentro de la IA, como la visión artificial o la robótica, pero en este tema nos centraremos en el aprendizaje automático y en la ingeniería del conocimiento.

El **aprendizaje automático** (AA) es la rama de la inteligencia artificial en la que se diseñan mecanismos para dotar a los sistemas computacionales de capacidad de aprendizaje, en el sentido de la capacidad de descubrir **regularidades (patrones)** en datos o situaciones anteriores y aplicarlos a nuevos problemas o situaciones análogos.

Por otro lado, la **ingeniería del conocimiento** (IC) es la encargada del desarrollo de sistemas basados en el conocimiento (SBC), como pueden ser los sistemas de ayuda a la decisión (SAD, *decision support systems* en inglés). La tradición de los SBC comenzó con los denominados **sistemas expertos**, sistemas computacionales que tratan de emular las capacidades de un experto en un tema basándose en la extracción del conocimiento del propio experto o de grupo de expertos y transmitiéndoselo al sistema. Con la proliferación del almacenamiento y el uso de datos de forma masiva, los SBC actuales suelen apoyarse en ambos pilares: expertos y datos.

Para la gestión del conocimiento experto hay diversas metodologías que consisten básicamente en la adquisición, la representación y la implantación de dicho conocimiento (IC). Estos sistemas suelen usar bases de reglas del tipo "Si el paciente tiene los síntomas A, B y C, entonces con **probabilidad o creencia X** tiene la enfermedad E" para almacenar y usar este conocimiento para inferir nuevos consejos de ayuda en la decisión.

Cuando se tienen en cuenta datos, por ejemplo, historias clínicas de los pacientes, casos anteriormente tratados, registros de incidencias de enfermedades, datos de factores que pueden provocar determinadas dolencias (factores medioambientales, hábitos sociales...), entonces es necesario recurrir a lo que en IA se llaman técnicas de aprendizaje automático y, por supuesto, técnicas provenientes de la matemática y la estadística, como las de **regresión** (formalización de una relación significativa entre dos o más variables para calcular pronósticos a partir del conocimiento de los valores en un individuo concreto).

Aunque muchas de estas técnicas son propias tanto de la estadística como del AA, se pueden considerar diversos paradigmas y grupos de técnicas dentro de la AA:

- **Paradigma analógico (aprendizaje por analogía)**

Se pretende encontrar una solución a un problema que se presenta ahora usando el mismo procedimiento utilizado en la resolución de uno similar que se presentó en otra ocasión anterior. Si dos problemas son similares en algún aspecto de su formulación, entonces pueden serlo también en sus soluciones. Se pueden abordar nuevos problemas reduciéndolos a problemas análogos resueltos. Ejemplos: analogía por transformación, analogía por derivación, razonamiento basado en casos, etc.

- **Paradigma inductivo**

Árboles de decisión, algoritmos de inducción pura, etc.

- **Paradigma conexionista**

Redes neuronales artificiales, etc.

- **Paradigma evolutivo**

Algoritmos genéticos, otros métodos de optimización, colonias de insectos, descenso estocástico del gradiente, etc.

- **Modelos gráficos probabilistas**

Bayesianos, cadenas de Markov, filtros de Kalman, redes de creencia, máquinas de soporte vectorial, metaheurísticas, etc.

Las técnicas de **agrupamiento** consisten en agrupar los elementos de una colección en subconjuntos (clases, categorías y **clusters**), nítidos o borrosos, en función de su similitud. Se trata de aprendizaje **no supervisado** porque las clases o categorías no se conocen *a priori*, sino que las determinan las propias similitudes entre los elementos. Por lo tanto, se centran en una medida de similitud entre elementos, de la que puede haber infinidad de variantes: estadísticas, distancias euclídeas, distancias vectoriales (coseno), distancias borrosas, etc. Veamos algunos ejemplos:

- **Paradigma conexionista (redes neuronales artificiales)**

Mapas de Kohonen, mapas SOM (*self organized maps* en inglés), como el SOM Toolbox de Matlab, etc.

- **Modelos estadísticos y probabilistas**

K-means, c-means, K-nearest neighbours (KNN), mean shift (ventanas circulares con un centroide), *dirichlet process* (estocásticos basados en distribuciones de probabilidad), *latent dirichlet allocation (LDA)*, modelos gaussianos, etc.

- **Extensiones basadas en lógica borrosa**

Fuzzy K-means, Fuzzy c-means, isodata, etc.

Las técnicas de **clasificación** (aprendizaje **supervisado**) consisten en asignar una clase a un nuevo elemento a partir de un conjunto de categorías previamente establecidas, por ejemplo, evaluar los síntomas de un nuevo paciente y decir que tiene gripe (clase previamente establecida). Se basan en un entrenamiento a partir de ejemplos con la solución conocida para crear modelos que permitan clasificar nuevos casos análogos:

- **Paradigma inductivo (árboles de decisión)**

ID3, CART, C4.5, See5, *random forest* (de moda en *big data*, introducidos por Leo Breiman), etc.

- **Paradigma conexionista (redes neuronales artificiales)**

Perceptrón multicapa (con *backpropagation*), redes convolutivas, neocognitrones, redes de Hopfield, redes recurrentes, redes *adaline, deep learning* (de moda en *big data*), etc.

- **Modelos estadísticos y probabilistas**

Redes bayesianas, clasificadores Bayes naíf, máquinas de soporte vectorial (SVM), metaheurísticas, etc.

3.8.4. Adecuación de los métodos a los problemas

La **analítica descriptiva** está orientada a la generación de un resumen claro y fácil de entender de una colección de datos. Este es el fundamento y el concepto más básico de todas las estadísticas. Se centra en la **visualización** de datos para entender el pasado y el presente. Se describen los datos con tablas o gráficos, mostrando descripciones numéricas de la variabilidad y la posición. También se suele llamar **modelización descriptiva**.

Para el **análisis predictivo** se suele usar:

- **Extrapolación** de funciones (tendencia para el futuro, pero sin capacidad de pronóstico, sobre hechos o cambios puntuales).
- **Correlaciones** entre variables (demasiado evidentes, no suelen funcionar de forma muy fina).
- Búsqueda de **patrones** en los datos que puedan ser aplicados a situaciones futuras (**KDD** y minería de datos).
- Métodos de **agrupamiento y clasificación**.
- **Extrapolación** de funciones (por ejemplo, estimaciones o líneas de tendencia).

Dentro del análisis predictivo, desempeña un papel fundamental el **análisis de series temporales**. Se suelen clasificar en:

- **Estacionarias** (las medias y/o la variabilidad se mantienen constantes).
- **No estacionarias** (las medias y/o la variabilidad no se mantienen constantes, sino que hay cambios de varianza o tendencias).

También es importante señalar qué otros métodos se usan para otras diferentes necesidades en el análisis de series temporales:

- **Tendencias**: método de mínimos cuadrados. Tendencias evolutivas y diferenciación estacional.
- **Predicción**: alisadores exponenciales:
 - Alisado exponencial simple.
 - Alisado exponencial lineal de Holt.
 - Alisado exponencial estacional de Holt-Winters.
- **Interpolación**: predicción de datos que faltan.

Mientras que el análisis predictivo se centra en un escenario futuro, el **prescriptivo** se centra en múltiples alternativas. Por lo tanto, un modelo prescriptivo puede ser considerado como una combinación de modelos predictivos (uno por cada posible escenario) que se ejecutan en paralelo.

El objetivo es encontrar la mejor opción posible: la **optimización**. Las principales técnicas usadas son las siguientes:

- Técnicas de investigación operativa.
 - Algoritmos genéticos.
 - Técnicas estocásticas.
 - Metaheurísticas.

A continuación se presentan las ventajas competitivas de los diferentes tipos de analítica:

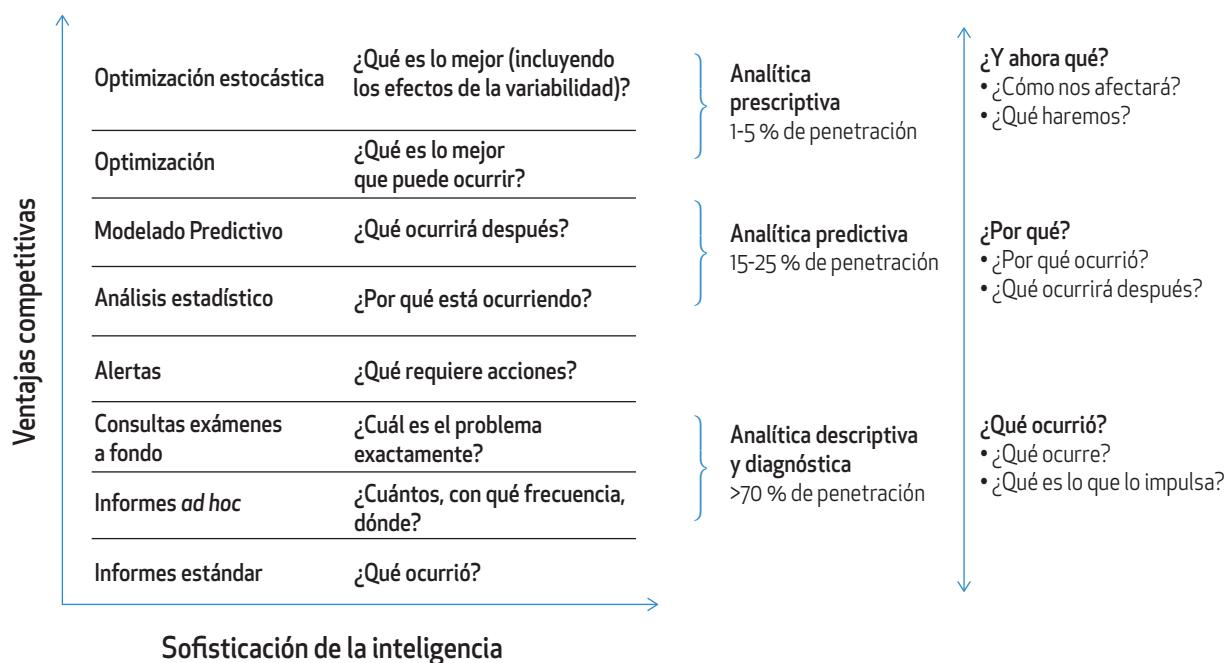


Figura 12. Ventajas competitivas de los diferentes tipos de analítica. Adaptado de *Competing on Analytics: The New Science of Winning*, por T. H. Davenport y J. G. Harris, 2007, Harvard: Harvard Business Press.



Capítulo 4

Algunas aplicaciones y ejemplos

A continuación se muestran varios ejemplos de aplicación de técnicas de razonamiento aproximado.

4.1. Prevención de incendios forestales basada en prototipos deformables borrosos

Se describe un proceso completo de KDD-minería de datos, presentado inicialmente en la tesis doctoral del autor de este manual (Olivas, 2000). Se parte de una base de datos de en torno a 100.000 registros con unos 40 campos que contienen datos sobre cada uno de los incendios forestales producidos en Galicia en los años anteriores a la fecha de desarrollo del trabajo. En particular se considerarán unos 12.000 registros de los años 1991 y 1992 (están los de todos los años).

4.1.1. Adquisición de conocimiento y datos

El proceso de adquisición del conocimiento es fundamental para la validez real de los métodos que pretenden servir para la resolución de problemas de ayuda en la toma de decisiones de este tipo. Aunque se describen con precisión estos procesos a lo largo de los diferentes apartados, parece adecuado abordar previamente ciertos aspectos.

1. Normalmente, la aparición de incendios forestales en la comunidad gallega está ligada a **factores socioeconómicos**, difíciles de cuantificar de una forma precisa, que no se reflejan en los datos estadísticos, ni en los informes o partes.

- 2.** La medida de factores físicos o meteorológicos no es lo precisa que podría ser deseable. Hay demasiados microclimas, especies arbóreas en poca superficie, pequeños cultivos, etc., como para que características como la humedad relativa o el estrés hídrico puedan ser generalizadas a grandes zonas de una forma fiable.
- 3.** En los datos estadísticos disponibles, existe un sesgo fundamental: no reflejan la auténtica peligrosidad e importancia de los incendios ocurridos.

Esto es debido a lo siguiente: supóngase la aparición de dos alarmas simultáneas en una misma comarca. Una de ellas es en una zona de matorrales que habitualmente se utiliza para pastoreo y en la que periódicamente se realizan quemas controladas, sin riesgo por proximidad a una zona de especial valor ecológico, y la otra alarma surge a pocos metros de un bosque con difícil accesibilidad. Es razonable y habitual, en el caso de que los recursos sean limitados, destinar más medios al segundo caso que al primero, lo que provoca en muchas ocasiones que en las estadísticas posteriores se refleje como importante el primer incendio (por ejemplo, 200 ha de matorral), y el segundo incluso ni aparezca (el umbral de aparición en estadísticas fue de 0,5 ha y posteriormente de 0,1 ha) porque se ha atajado en sus inicios.

- 4.** Los criterios de prioridad, por ejemplo, para el envío de medios aéreos, no siempre están claros, porque, en los puntos donde se toman estas decisiones, la información de la que se dispone no es lo suficientemente objetiva ni completa.

Teniendo en cuenta estos criterios, se ha realizado un análisis previo y superficial de todos los ciclos de incendios de una comarca durante varios años:

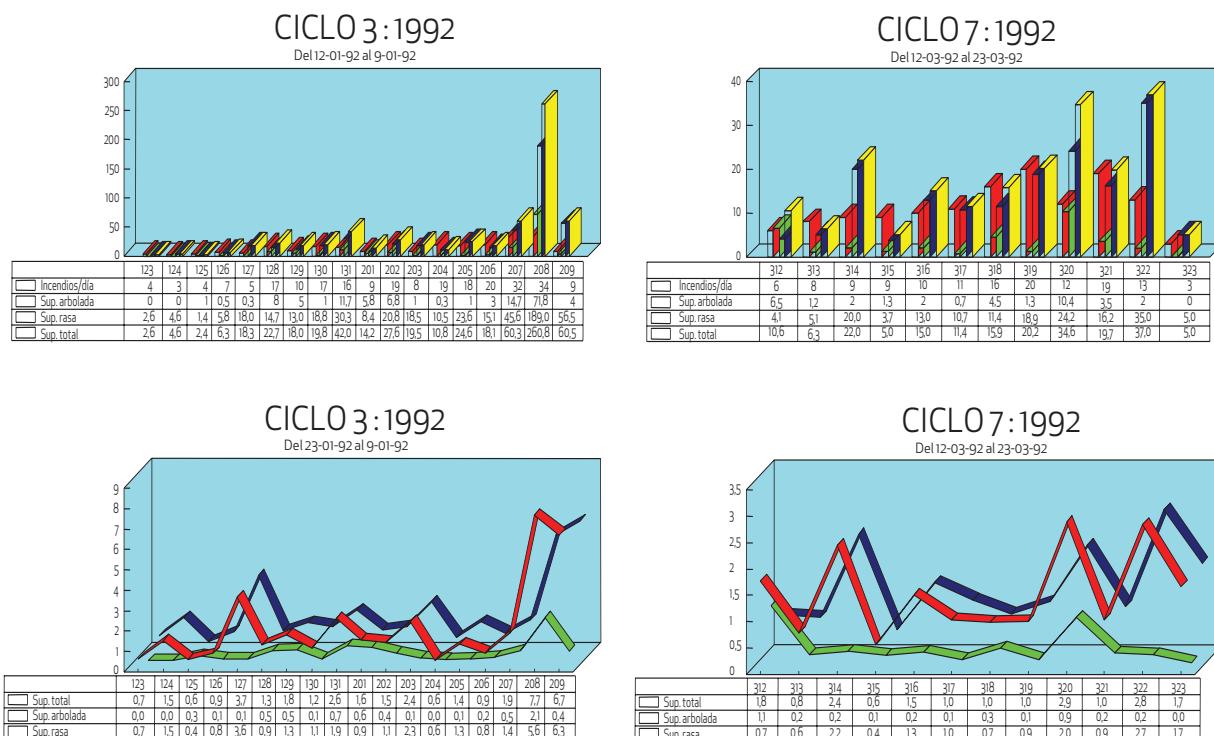


Figura 13. Análisis previo, prospección para verificar la viabilidad de las hipótesis iniciales.

Se ha observado que la evolución de la siniestralidad puede ser representada como una función creciente de tipo sigmoidal, dividida en tres sectores y con inicio en el día posterior a un periodo de lluvia en el que el número de los incendios se ha reducido a cero.

Este patrón de crecimiento se repite de forma cíclica después de cada periodo de este tipo, pero puede sufrir modificaciones debido a factores específicos. La Figura 14 muestra la representación del patrón de crecimiento:

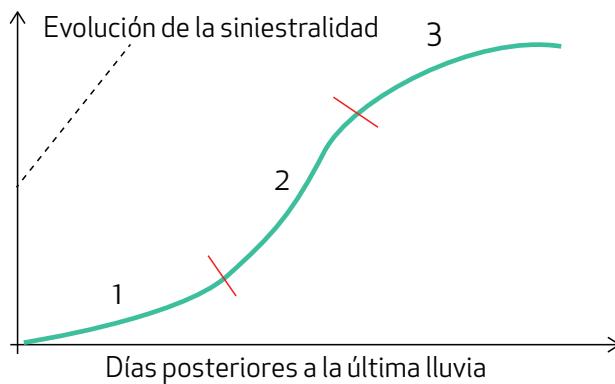


Figura 14. Representación de un posible patrón de evolución de incendios forestales en Galicia.

Entiéndase por siniestralidad la combinación de varios factores, como número, peligrosidad de los incendios, etc.

- El primer sector representa un lento crecimiento de la siniestralidad en los días inmediatamente posteriores al periodo de lluvias.
- El segundo sector expresa un alto crecimiento de la siniestralidad, especialmente en cuanto al número de incendios.
- En el tercero se quiere representar una estabilización en cuanto al número, pero un progresivo aumento en la peligrosidad de los fuegos.



En nuestra vida cotidiana, es usual asociar un hecho o conjunto de hechos con un **patrón aprendido**, de tal forma que el patrón interpreta la situación y de él dependen las acciones que llevemos a cabo. Por ejemplo, si estamos conduciendo y empieza a granizar adecuamos nuestra conducción al esquema de conducción bajo condiciones potencialmente peligrosas que nuestra experiencia ha forjado en nuestro conocimiento.

Como se ha visto en el ejemplo, muchas de las acciones que realizamos en nuestra vida dependen de una interpretación. Lo que aquí se plantea es que interpretar una situación es encontrar en los datos los **patrones o prototipos afines** a las circunstancias del problema. En este caso, se trata de simular la capacidad del experto para interpretar la situación, es decir, para encontrar el modelo de evolución de la siniestralidad de los incendios más adaptado a las circunstancias reales.

4.1.2. Concepto y prototípos (la importancia del científico de datos)



La importancia del científico de datos se pone de manifiesto cuando es necesaria (casi siempre) una visión multidisciplinar del análisis del problema, buscando, en muchos casos, **soluciones sofisticadas** (desde el punto de vista de la inteligencia artificial), porque con el uso de herramientas y técnicas estándar simples la aproximación a muchos problemas de análisis de datos reales puede ser trivial y no proporcionar resultados relevantes.

Por ejemplo, tomando como marco de referencia la **teoría de prototipos** de la **psicología cognitiva**, podría entenderse que esta representación es prototípica del avance de la siniestralidad de los incendios. Sin embargo, en el proceso de adquisición del conocimiento, se pudo observar que esta representación simplifica en exceso las pautas del comportamiento de los expertos. Cuando un técnico se enfrenta a una situación real, maneja un abanico de prototipos determinados por una serie de factores, es decir, debe decidir qué tipo de evolución de la siniestralidad es previsible. Dicho de otro modo, el prototipo de la **evolución de la siniestralidad** no es único, sino que existen diferentes formas de evolución dentro de la misma estructura sigmoidal.

El profesor Lotfi A. Zadeh, creador de la **lógica borrosa**, aludía a las **teorías clásicas de prototipos** desde el punto de vista de la psicología (Zadeh, 1982), criticando precisamente lo que aquí se ha expuesto: su falta de adecuación a la función que debe cumplir un prototipo. La aproximación de Zadeh a lo que debe entenderse por prototipo es menos intuitiva que las concepciones de las teorías psicológicas, pero más racional, más próxima a lo que en un examen detenido muestra el significado de un concepto prototípico.



Enlace de interés

En el blog de la VIU se puede leer el artículo "Zadeh (DEP), la lógica borrosa y el análisis de datos masivos":

<https://www.universidadviu.es/zadeh-d-e-p-la-logica-borrosa-analisis-datos-masivos/>

En nuestro caso, se ha observado que la idea de Zadeh sugiere que un concepto engloba un conjunto de prototipos, los cuales representan la buena, baja o media compatibilidad de los ejemplares con el concepto. Las **categorías prototípicas borrosas** representan las diferentes clases que se pueden determinar en el dominio.

Desde este punto de vista, se puede hablar de siniestralidad altamente progresiva, medianamente progresiva o escasamente progresiva.

Esto se puede representar, tal y como refleja la Figura 15, con tres sigmoides, entendidas simplemente como representaciones gráficas de los tres prototipos borrosos:

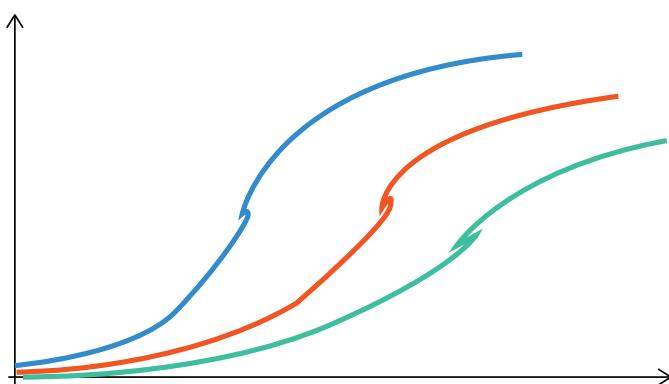


Figura 15. Representación de los tres patrones de evolución.

4.1.3. Descubrimiento de conocimiento prototípico borroso en datos sobre incendios forestales

El proceso clásico de KDD, aquí orientado al **descubrimiento prototípico** para este caso particular de los datos disponibles sobre incendios forestales se lleva a cabo en las siguientes fases:

1. **Selección.** Aplicando el conocimiento del dominio y el conocimiento relevante a priori, teniendo en cuenta los objetivos del proceso global de KDD, se crea una tarjeta de datos (datos objetivo) que incluye conjuntos seleccionados de datos o subconjuntos de variables relevantes o ejemplos. Se toma como conjunto de partida una base de datos relacional que contiene aproximadamente 12.000 incendios ocurridos en Galicia durante los años 1991 y 1992. Se seleccionan estos años por ser 1991 un año poco conflictivo y 1992 uno muy conflictivo. Se seleccionan los 3.204 correspondientes a la comarca de Lugo, compuesta por siete municipios, de los años 1991 y 1992. También se separan por comarcas y se eliminan campos no relevantes.
2. **Preproceso:** limpieza de datos, eliminación de ruidos, manejo de campos vacíos, datos perdidos, valores desconocidos o por defecto, evolución de datos. Se aplican técnicas estándar de bases de datos.
3. **Transformación:** reducción del número de variables, localización de ciclos de incendios. Se buscan formas útiles para expresar los datos dependiendo del uso posterior que se les va a dar y de los objetivos del sistema. Se usa el conocimiento experto, técnicas de transformación e informes en bases de datos. Se ordenan y clasifican los ciclos según la evolución de la siniestralidad.
4. **Data mining:** elección de los algoritmos. Se decide el modelo que se deriva del algoritmo de *data mining* elegido (clasificación, resumen de datos, predicción). Se buscan patrones de interés, en cuanto a clasificación, reglas o árboles, regresión, clasificación, dependencia, heurísticas e incertidumbre. Se generan los prototipos de evolución de la siniestralidad a partir de los ciclos representativos de cada una de las clases.

A continuación se muestran con detalle las operaciones de preproceso, transformación (clustering jerárquico y ordenación de los ciclos) y *data mining*.

Preproceso, eliminación de ruido

Una vez extraídos de la base de datos de Galicia los 3.204 incendios correspondientes a los años 1991 y 1992 de la comarca de Lugo, se debe proceder a su estudio completo. Antes habrá que hacer unas modificaciones tanto en el diseño como en el contenido de la tabla, ya que presenta diversas irregularidades que hacen imposible o muy dificultoso el aprovechamiento de los datos.

- Tratamiento de las fechas y horas

La base de datos contiene el control de las fechas y las horas en que los incendios se han iniciado, se han controlado y se han extinguido. Estos datos son muy importantes, ya que permiten el control de la evolución temporal de los incendios y la obtención de datos estadísticos sobre cada día, semanas... En la base de datos inicial, estos campos estaban tipificados como simples **cadenas de caracteres con un cierto formato**, lo cual facilitaba en un primer momento su almacenamiento, pero dificulta enormemente su posterior tratamiento, ya que pierden toda su semántica propia, no pudiéndose efectuar sobre ellos operaciones como comparaciones, restas o distancias.

La solución a este problema es sencilla, siempre y cuando **las fechas y horas hayan sido almacenadas con algún formato estándar**, lo que en este caso ocurre. De esta forma, cambiando el tipo de los campos en el sistema gestor de base de datos utilizado (Microsoft Access '97) se permite que las aplicaciones accedan y operen sobre estos datos con toda la semántica correspondiente a los tipos fecha y hora (CTime en Visual C++):

Tabla 2
Modificación de fechas y horas

Campo afectado	Formato	Descripción
FECHA_INI	DD/MM/AA	Fecha de inicio del incendio.
HORA_INI	HH:MM:SS	Hora de inicio del incendio.
FECHA_EXT	DD/MM/AA	Fecha de control del incendio.
HORA_EXT	HH:MM:SS	Hora de control del incendio.
FECHA_FIN	DD/MM/AA	Fecha de la extinción total.
HORA_FIN	HH:MM:SS	Hora de la extinción total.

- Pérdida de datos numéricos

Todos los campos o variables estadísticas que se manejan poseen un conjunto de valores perdidos, es decir, el valor o **los valores que no se consideran como válidos para la variable** con la que se está trabajando, no existiendo por defecto valores ausentes.

Todos los valores numéricos existentes en esta base de datos son indispensables para el posterior tratamiento. Por ello es necesario que los valores perdidos estén bien definidos y se puedan procesar de forma correcta.

Existen dos tipos de valores perdidos, los **valores omitidos por el usuario** (*missing values* en inglés), códigos que indican que el verdadero valor de una variable es desconocido y que los casos que contengan esos valores deben ser excluidos del análisis, y los **valores perdidos por el sistema** (*system-missing values* en inglés), valores asignados por el SGBD correspondiente cuando un valor de los datos resulta indefinido de acuerdo con el tipo de formato que se ha especificado (como un valor identificativo de menos infinito).

En la tabla de incendios que se maneja, y debido en su mayor parte a su procedencia, sucede que los valores numéricos han sido introducidos a mano y sin tener en cuenta que debían tener un valor por defecto. Así pues, en la primera parte de la base de datos ocurre que todos los campos están totalmente ocupados por valores que suelen tomar el valor 0 (superficie quemada despreciable, categorías de personal que no estuvieron en el incendio...). El problema surge cuando alrededor de la mitad de la tabla esos valores no aparecen y campos numéricos que deberían tener algún valor no lo tienen, lo cual provoca un cierto caos en su tratamiento, ya que el SGBD utilizado y el controlador de ODBC para comunicarlo con la aplicación colocan en estos valores el de menos infinito ($-9,18 \times 10^{-19}$), lo cual hace imposible un cálculo automático sobre estos valores si no se controla esta contrariedad:

Tabla 3

Preparación de algunos datos numéricos

Campo afectado	(Min, Max)	Tipo de dato	Descripción
SUP_ARBO	(0, 30)	real doble	Superficie arbolada arrasada por el incendio
SUP_RASA	(0, 60)	real doble	Superficie rasa afectada por el incendio
SUP_TOTAL	(0, 90)	real doble	Superficie total quemada por el incendio
TECNICOS	(0, 10)	entero	Núm. de técnicos que intervinieron en el incendio
AGEN_FOR	(0, 10)	entero	Núm. de agentes forestales implicados
AUT_CIVIL	(0, 15)	entero	Autoridad civil destinada al incendio

Debido a la experiencia acumulada en la gestión de la información que generan los incendios, se sabe que estos valores son nulos, despreciables o desconocidos por su poca relevancia. Por eso, se ha asignado a todos estos casos un valor por defecto 0.

- Malos diseños

El diseño de una base de datos que gestiona la información sobre un cierto asunto (en este caso incendios) debe ser cuidadoso, ya que posteriormente se debe acceder a ellos para obtener resultados de todo tipo. Esta es una premisa que se aleja mucho de la realidad de la base de datos con la que aquí se trabaja.

Esta base de datos de una única tabla que gestiona todos los datos necesarios está concebida por su fácil construcción y modificación, sin tener en cuenta para nada los conceptos de modelización y normalización de las bases de datos ni las características esenciales de un buen diseño. Este diseño está más destinado a almacenar de alguna forma sencilla los datos sin ningún interés en una posterior utilización de estos.

Estas situaciones son la pesadilla de las personas dedicadas a la analítica de datos, ya que no se trata de valores perdidos, sino de valores que existen sin saberse dónde se localizan.

Esta situación se presenta, en nuestro caso, cuando se quieren reflejar los recursos personales y mecánicos que estuvieron involucrados en la extinción de cierto incendio:

- Para medios humanos: 20 campos de tipo texto con los identificadores de cada uno de los recursos. Solo se ocupan los campos, por orden, según el número de recursos asignados; el resto de los campos hasta 20 se quedan en blanco. De estos 20 campos, solo se utilizan con frecuencia los 5 primeros, y como máximo se utilizan 14, por lo que hay campos que sobran.
- Para medios mecánicos: de los 20 campos de idénticas características que los anteriores, solo se utilizan alguna vez 10, de los cuales solo los primeros son utilizados con frecuencia.

El diseño, aunque posee un cierto margen para poder albergar muchos recursos asignados a incendios, genera espacio desaprovechado. Si se produjera una situación con más de 20 recursos (humanos o mecánicos), la base de datos no podría contemplarlos. Además, para obtener datos estadísticos directamente de la base de datos sobre el personal o vehículos utilizados para la obtención del incendio (dato muy relevante), hay que hacer comprobaciones que generarían una pérdida de tiempo y quizás también de información que no sería aceptable en un estudio serio de los incendios.

Para no tener que hacer cálculos inútiles e imprecisos sobre estos datos, se han creado dos campos nuevos en la tabla: personal y medios, los cuáles son, respectivamente, un recuento de los medios humanos y mecánicos que han sido asignados al incendio. Estos campos han sido llenados por un algoritmo que ha contado los campos no vacíos de los indicados anteriormente ofreciendo un resultado acumulado de todos ellos. Posteriormente se han eliminado todos los 40 campos que la base de datos destinaba a indicar el personal y los medios.

Los únicos datos perdidos en este proceso son los identificadores correspondientes a los recursos existentes en la comarca, datos que pueden ser obtenidos fácilmente de otras fuentes. Como contrapartida de 40 campos prácticamente inútiles, se ha pasado a dos campos numéricos, sin valores perdidos, que ofrecen valores muy significativos para el posterior estudio de estos datos. Así pues, se obtienen unos datos más manejables y más significativos:

Tabla 4

Resumen de los campos de personal y medio

Campos introducidos	Descripción
PERSONAL	Cantidad de personas (medios humanos) no especialistas destinadas al incendio.
MEDIOS	Cantidad de medios mecánicos (motobombas, palas...) destinados al incendio.
Campos eliminados	Descripción
De PER1 a PER20	Campos destinados a los identificadores de los recursos humanos involucrados en la extinción del incendio.
De MED1 a MED20	Campos destinados a los identificadores de los recursos mecánicos involucrados en la extinción del incendio.

Para todas las labores que siguen, se ha construido una aplicación que permite que todas las operaciones se realicen paso a paso, comprobando cada vez los efectos sobre la base de datos, y albergando todos los algoritmos y funciones auxiliares indicadas durante la exposición del preproceso y la primera transformación de los datos.

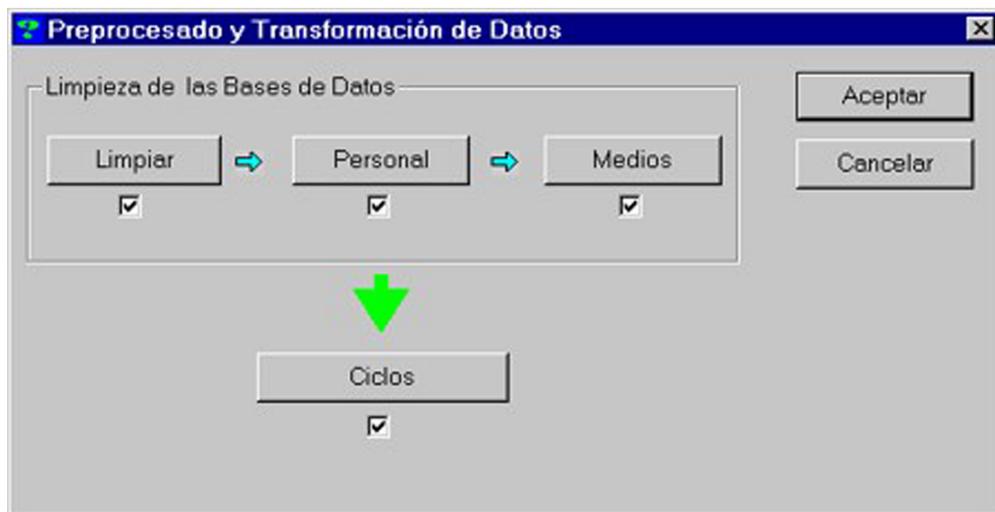


Figura 16. Interfaz de las operaciones descritas.

En esta interfaz se pueden ver los diferentes botones que suponen la utilización de cada uno de los algoritmos sobre la base de datos y un control *checkbox* para indicar que cada proceso se ha culminado con éxito.

El resultado de este proceso han sido 37 ciclos que quedan reflejados en la siguiente tabla:

Tabla 5

Ciclos de incendios 1991-1992

A	B	C	D	E	F	G	H	I	J	K	L
0	2/01/91	2/01/91	1	0	0	1	0,01	0,5	3	1	1
1	15/01/91	15/01/91	1	1	0	0	0	0,001	1	1	0
2	24/01/91	30/01/91	9	2	4	3	0	0,272	10	10	2
3	23/02/91	25/02/91	12	1	8	3	0	0,395	14	13	2
4	4/03/91	4/03/91	1	0	0	1	0,008	0,03	3	1	1
5	11/03/91	11/03/91	1	0	1	0	0	0,03	1	1	0
6	20/03/91	20/03/91	3	1	0	2	0,032	0,13	2	3	1
7	29/03/91	2/04/91	45	17	19	9	0,213	1,209	31	60	45
8	9/04/91	11/04/91	20	6	10	4	0	0,406	20	23	14
9	14/04/91	24/04/91	48	24	14	10	0,392	0,824	64	71	40
10	27/04/91	27/04/91	1	1	0	0	0	0,01	0	1	1
11	8/05/91	29/05/91	113	59	18	36	0,99	3,641	162	185	126
12	1/06/91	8/06/91	7	4	2	1	0	0,076	9	7	5

>>>

Tabla 5

Continuación

A	B	C	D	E	F	G	H	I	J	K	L
13	11/06/91	19/06/91	19	15	0	4	0,169	0	24	30	30
14	22/06/91	6/07/91	57	39	10	8	0	1,841	72	102	87
15	10/07/91	10/09/91	554	375	94	85	0	14,922	744	1306	1196
16	14/09/91	25/09/91	87	58	24	5	0	1,335	94	196	198
17	2/10/91	2/10/91	1	1	0	0	0	0	0	1	0
18	6/10/91	7/10/91	8	6	2	0	0	0,109	10	20	6
19	1/11/91	1/11/91	1	0	1	0	0	0,03	1	2	0
20	24/11/91	24/11/91	1	0	1	0	0	0,02	4	2	1
21	29/11/91	30/11/91	2	1	1	0	0	0,03	2	2	0
22	5/12/91	7/12/91	16	10	5	1	0	0,396	9	25	7
23	14/12/91	16/12/91	6	1	4	1	0	0,245	7	14	3
24	21/12/91	8/01/92	213	98	78	37	63,876	434,756	147	383	136
25	16/01/92	20/01/92	9	8	1	0	1,01	2,71	7	12	4
26	23/01/92	11/02/92	424	233	112	79	206,3	1064,74	234	635	187
27	15/02/92	23/03/92	916	635	189	92	183,1	1074,52	366	1406	515
28	10/04/92	23/05/92	268	195	41	32	139,2	293,4	223	494	305
29	13/06/92	15/06/92	7	6	0	1	0,8	3,24	7	9	5
30	19/06/92	20/06/92	3	2	0	1	1,7	1,8	2	4	3
31	25/06/92	26/06/92	2	2	0	0	0	0,3	1	2	0
32	29/06/92	18/07/92	53	51	1	1	2,08	8,57	38	70	54
33	21/07/92	7/08/92	133	110	9	14	0	82,9	136	266	217
34	12/08/92	27/08/92	80	70	6	4	6,15	0	68	155	142
35	2/09/92	21/09/92	80	74	4	2	3,92	0	87	155	142
36	3/10/92	3/10/92	1	0	0	1	1	0,75	1	2	1

Leyenda:

- A:** Número de Identificación
- B:** Fecha de inicio del ciclo
- C:** Fecha de finalización del ciclo
- D:** Número total de sucesos
- E:** Número de conatos
- F:** Número de quemas
- G:** Número de incendios
- H:** Superficie arbolada quemada
- I:** Superficie rasa quemada
- J:** Número de especialistas (técnicos, agentes forestales...)
- K:** Número de brigadas de personal
- L:** Número de medios mecánicos (terrestres, aéreos...)

La Tabla 5 muestra los 37 ciclos de incendios entre dos períodos de lluvias que han tenido lugar en la zona estudiada durante los años 1991 y 1992. Los siguientes pasos serán comprobar qué estructura de clases determinan los propios datos de estos ciclos (agrupamiento o *clustering*), establecer una correlación entre los *clusters* descubiertos y los prototipos sugeridos por los expertos y, por último, encuadrar cada uno de estos ciclos en su correspondiente prototipo.

Es decir, primero se ha realizado un agrupamiento sensible al contexto (detección de ciclos) y a continuación se hará uno en un espacio lingüístico inducido (se comenzará con un proceso de agrupamiento sobre la tabla de ciclos y se concluirá con uno de clasificación de cada ciclo en un prototipo, tal como se había definido en las etapas del KDD, "convertir agrupamiento en clasificación...").

Agrupamiento jerárquico sobre la tabla de ciclos

Con el fin de detectar las relaciones entre los ciclos, para obtener aquellos de escasa, mediana y alta siniestralidad, se realiza un proceso de *clustering* jerarquizado mediante la técnica de emparrillados (*repertory grids* en inglés). El conjunto de elementos es el constituido por los 37 ciclos, y las construcciones son las 7 que se detallan a continuación:

Tabla 6
Construcciones utilizadas

	Construcción	Número de incendios, C1	
Valores	Muy pocos	[0-20]	1
	Pocos	[21-30]	2
	Regular	[31-50]	3
	Bastantes	[51-100]	4
	Muchos	[101-...]	5

	Construcción	Superficie arb. quemada, C2	
Valores	Muy poca	[0-0,2]	1
	Poca	[0,21-0,4]	2
	Regular	[0,41-1]	3
	Bastante	[1,1-20]	4
	Mucha	[21-...]	5

	Construcción	Superficie rasa quemada, C3	
Valores	Muy poca	[0-1]	1
	Poca	[1,1-5]	2
	Regular	[5,1-100]	3
	Bastante	[101-1000]	4
	Mucha	[1001-...]	5

>>>

Tabla 6*Continuación*

Construcción		Número de especialistas, C4	
Valores	Muy pocos	[0-10]	1
	Pocos	[11-20]	2
	Regular	[21-50]	3
	Bastantes	[51-100]	4
	Muchos	[101-...]	5

Construcción		Número de trabajadores, C5	
Valores	Muy pocos	[0-10]	1
	Pocos	[11-50]	2
	Regular	[51-100]	3
	Bastantes	[101-300]	4
	Muchos	[301-...]	5

Construcción		Núm. de medios, C6	
Valores	Muy pocos	[0-10]	1
	Pocos	[11-30]	2
	Regular	[31-50]	3
	Bastantes	[51-100]	4
	Muchos	[101-...]	5

Construcción		Núm. de días, C7	
Valores	Muy pocos	[0-3]	1
	Pocos	[4-10]	2
	Regular	[11-20]	3
	Bastantes	[21-30]	4
	Muchos	[31-...]	5

Así pues, la malla de repertorio queda como se refleja en la Tabla 7:

Tabla 7*Matriz de entrada al algoritmo de agrupamiento jerárquico*

Ciclo	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Ciclo
C1	1	1	1	1	1	1	1	3	1	3	1	5	1	1	4	5	4	1	1	C1
C2	1	1	1	1	1	1	1	2	1	2	1	3	1	1	1	1	1	1	1	C2
C3	1	1	1	1	1	1	1	2	1	1	1	2	1	1	2	3	2	1	1	C3
C4	1	1	1	2	1	1	1	3	2	4	1	5	1	3	4	5	4	1	1	C4
C5	1	1	1	2	1	1	1	3	2	3	1	4	1	2	4	5	4	1	2	C5
C6	1	1	1	1	1	1	1	3	2	3	1	5	1	2	4	5	5	1	1	C6
C7	1	1	2	1	1	1	1	2	1	3	1	4	2	2	3	5	3	1	1	C7

>>>

Tabla 7

Continuación

Ciclo	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	Ciclo
C1	1	1	1	1	1	5	1	5	5	5	1	1	1	4	5	4	4	1	C1
C2	1	1	1	1	1	5	4	5	5	5	3	4	1	4	1	4	4	3	C2
C3	1	1	1	1	1	4	2	5	5	4	2	2	1	3	3	1	1	1	C3
C4	1	1	1	1	1	5	1	5	5	5	1	1	1	3	5	4	4	1	C4
C5	1	1	1	2	2	5	2	5	5	5	1	1	1	3	4	4	4	1	C5
C6	1	1	1	1	1	5	1	5	5	5	1	1	1	4	5	5	5	1	C6
C7	1	1	1	1	1	3	2	3	5	5	1	1	1	3	3	3	3	1	C7

Para realizar un análisis de *clusters* (*clustering jerárquico*) sobre elementos, se construye una **matriz de proximidad** (malla de repertorio reducida) que representa las diferentes similitudes de los elementos, una matriz de 37×37 elementos (los ciclos) que por encima de la diagonal representan las distancias entre los diferentes ciclos. Pasando estos valores a porcentajes y se crea la tabla reducida a porcentaje.

Como resultado se obtienen los siguientes porcentajes de similitud entre elementos y sus agrupaciones:

$(0, 1) \rightarrow 100\%$
 $((0, 1), 4) \rightarrow 100\%$
 $((((0, 1), 4), 5) \rightarrow 100\%$
 $(((((0, 1), 4), 5), 6) \rightarrow 100\%$
 $((((((0, 1), 4), 5), 6), 10) \rightarrow 100\%$
 $((((((0, 1), 4), 5), 6), 10), 17) \rightarrow 100\%$
 $((((((0, 1), 4), 5), 6), 10), 17), 19) \rightarrow 100\%$
 $((((((0, 1), 4), 5), 6), 10), 17), 19), 20) \rightarrow 100\%$
 $((((((0, 1), 4), 5), 6), 10), 17), 19), 20), 21) \rightarrow 100\%$
 $((((((0, 1), 4), 5), 6), 10), 17), 19), 20), 21), 31) \rightarrow 100\% (A)$

$(2, 12) \rightarrow 100\% (B)$

$(18, 22) \rightarrow 100\%$
 $((18, 22), 23) \rightarrow 100\% (C)$

$(34, 35) \rightarrow 100\% (D)$

$((((((0, 1), 4), 5), 6), 10), 17), 19), 20), 21), 31), (2, 12)) \rightarrow 97\%$
 $(3, 8) \rightarrow 97\%$
 $(14, 16) \rightarrow 97\%$
 $(24, 26) \rightarrow 97\%$
 $(27, 28) \rightarrow 97\%$
 $(29, 30) \rightarrow 97\%$

$((((((((0, 1), 4), 5), 6), 10), 17), 19), 20), 21), 31), (2, 12)), ((18, 22), 23)) \rightarrow 93\%$

$((29, 30), 36) \rightarrow 93\%$

$((((((((0, 1), 4), 5), 6), 10), 17), 19), 20), 21), 31), (2, 12)), ((18, 22), 23)), (3, 8)) \rightarrow 90\%$

$(7, 9) \rightarrow 90\%$

$(15, 33) \rightarrow 90\%$

$((24, 26), (27, 28)) \rightarrow 90\%$

$(25, ((29, 30), 36)) \rightarrow 90\%$

$(((((((((0, 1), 4), 5), 6), 10), 17), 19), 20), 21), 31), (2, 12)), ((18, 22), 23)), (3, 8)), 13) \rightarrow 83\%$

$(11, (15, 33)) \rightarrow 83\%$

$((14, 16), (34, 35)) \rightarrow 83\%$

$((7, 9), ((14, 16), (34, 35))) \rightarrow 79\%$

$(((((((((0, 1), 4), 5), 6), 10), 17), 19), 20), 21), 31), (2, 12)), ((18, 22), 23)), (3, 8)), 13), (25, ((29, 30), 36))) \rightarrow 75\%$

$((7, 9), ((14, 16), (34, 35))), 32) \rightarrow 75\%$

$((11, (15, 33)), ((24, 26), (27, 28))) \rightarrow 75\%$

$((((7, 9), ((14, 16), (34, 35))), 32), ((11, (15, 33)), ((24, 26), (27, 28)))) \rightarrow 61\%$

$(((((((((0, 1), 4), 5), 6), 10), 17), 19), 20), 21), 31), (2, 12)), ((18, 22), 23)), (3, 8)), 13), (25, ((29, 30), 36))), (((7, 9), ((14, 16), (34, 35))), 32), ((11, (15, 33)), ((24, 26), (27, 28)))) \rightarrow 40\%$

Así pues, se obtiene el siguiente dendrograma:

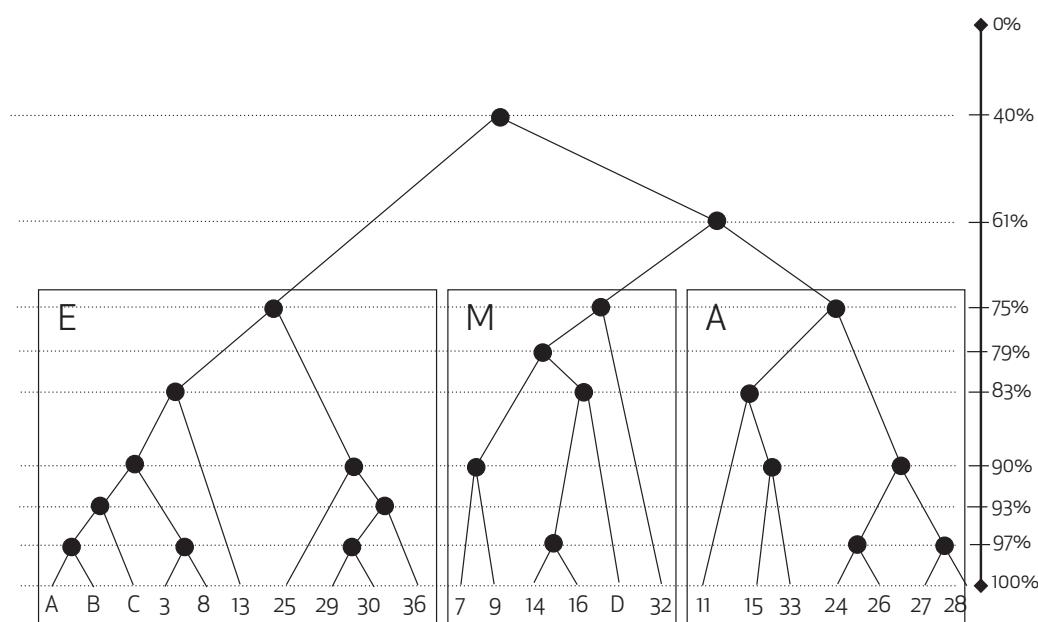


Figura 17. Dendrograma final del proceso de agrupamiento jerárquico.

Leyenda:

E: Siniestralidad escasamente progresiva. **M:** Siniestralidad medianamente progresiva. **A:** Siniestralidad altamente progresiva

Ordenación de los ciclos según la evolución de su siniestralidad

Una vez realizado este proceso de *clustering* o agrupamiento sobre los ciclos, se está en condiciones de ordenarlos en función de su siniestralidad. Para ello, realizados varios test a expertos, se ha llegado a la conclusión de que la evolución de la siniestralidad está básicamente vinculada al número de siniestros ocurridos (sin distinguir entre incendios, superficie arbolada mayor a media hectárea, conatos, superficie total menor a media hectárea y quemas, superficie rasa mayor que media hectárea), porque cualquiera de los siniestros, por pequeño que haya sido, podría haber tenido consecuencias peores de no haberse actuado a tiempo. El otro factor influyente en la evolución de la siniestralidad es la superficie que se ha quemado, dando un mayor peso a la superficie arbolada.

Teniendo en cuenta estos factores, se está en condiciones de definir una **medida heurística** de siniestralidad del ciclo, expresada del siguiente modo:

$$\begin{aligned} & \text{Número total de incendios / 100} \\ & + \\ & [(1 \times \text{superficie arbolada total}) + (0,5 \times \text{superficie rasa total})] / 100 \end{aligned}$$

Presentados casos de ejemplo, los expertos concuerdan en que un ciclo sería de siniestralidad **escaladamente** progresiva cuando esta medida estuviera por debajo de 0,4, de siniestralidad **medianamente** progresiva entre este valor y 1, y de siniestralidad **altamente** progresiva a partir de 1.

Tabla 8

Ordenación de los ciclos según su peligrosidad

A	B	C	D	E	F	G	H	I	J	K	L	M	N	P
17	2/10/91	2/10/91	1	1	0	0	0	0	0	1	0	1	0,01	E
1	15/01/91	15/01/91	1	1	0	0	0	0,001	1	1	0	1	0,0105	E
10	27/04/91	27/04/91	1	1	0	0	0	0,01	0	1	1	1	0,0105	E
20	24/11/91	24/11/91	1	0	1	0	0	0,02	4	2	1	1	0,0101	E
5	11/03/91	11/03/91	1	0	1	0	0	0,03	1	1	0	1	0,0105	E
19	1/11/91	1/11/91	1	0	1	0	0	0,03	1	2	0	1	0,0105	E
4	4/03/91	4/03/91	1	0	0	1	0,008	0,03	3	1	1	1	0,0103	E
0	2/01/91	2/01/91	1	0	0	1	0,01	0,5	3	1	1	1	0,0126	E
21	29/11/91	30/11/91	2	1	1	0	0	0,03	2	2	0	2	0,0205	E
31	25/06/92	26/06/92	2	2	0	0	0	0,3	1	2	0	2	0,0215	E
36	3/10/92	3/10/92	1	0	0	1	1	0,75	1	2	1	1	0,0235	E
6	20/03/91	20/03/91	3	1	0	2	0,032	0,13	2	3	1	1	0,0307	E
30	19/06/92	20/06/92	3	2	0	1	1,7	1,8	2	4	3	2	0,056	E
23	14/12/91	16/12/91	6	1	4	1	0	0,245	7	14	3	3	0,0625	E
12	1/06/91	8/06/91	7	4	2	1	0	0,076	9	7	5	8	0,0708	E
18	6/10/91	7/10/91	8	6	2	0	0	0,109	10	20	6	2	0,0805	E
2	24/01/91	30/01/91	9	2	4	3	0	0,272	10	10	2	7	0,0913	E
29	13/06/92	15/06/92	7	6	0	1	0,8	3,24	7	9	5	3	0,0942	E
25	16/01/92	20/01/92	9	8	1	0	1,01	2,71	7	12	4	5	0,1136	E

>>>

Tabla 8

Continuación

A	B	C	D	E	F	G	H	I	J	K	L	M	N	P
3	23/02/91	25/02/91	12	1	8	3	0	0,395	14	13	2	3	0,1219	E
22	5/12/91	7/12/91	16	10	5	1	0	0,396	9	25	7	3	0,1619	E
13	11/06/91	19/06/91	19	15	0	4	0,169	0	24	30	30	9	0,1916	E
8	9/04/91	11/04/91	20	6	10	4	0	0,406	20	23	14	3	0,2020	E
7	29/03/91	2/04/91	45	17	19	9	0,213	1,209	31	60	45	4	0,4581	M
9	14/04/91	24/04/91	48	24	14	10	0,392	0,824	64	71	40	11	0,4880	M
14	22/06/91	6/07/91	57	39	10	8	0	1,841	72	102	87	15	0,5792	M
32	29/06/92	18/07/92	53	51	1	1	2,08	8,57	38	70	54	20	0,5936	M
35	2/09/92	21/09/92	80	74	4	2	3,92	0	87	155	142	20	0,8392	M
34	12/08/92	27/08/92	80	70	6	4	6,15	0	68	155	142	16	0,8615	M
16	14/09/91	25/09/91	87	58	24	5	0	1,335	94	196	198	12	0,8766	M
11	8/05/91	29/05/91	113	59	18	36	0,99	3,641	162	185	126	22	1,1581	A
33	21/07/92	7/08/92	133	110	9	14	0	82,9	136	266	217	17	1,7445	A
24	21/12/91	8/01/92	213	98	78	37	63,876	434,756	147	383	136	18	4,9425	A
28	10/04/92	23/05/92	268	195	41	32	139,2	293,4	223	494	305	44	5,539	A
15	10/07/91	10/09/91	554	375	94	85	0	14,922	744	1306	1196	61	5,6146	A
26	23/01/92	11/02/92	424	233	112	79	206,3	1064,74	234	635	187	19	11,626	A
27	15/02/92	23/03/92	916	635	189	92	183,1	1074,52	366	1406	515	39	16,363	A

Leyenda:

- A:** Número de identificación
- B:** Fecha de inicio del ciclo
- C:** Fecha de finalización del ciclo
- D:** Número total de sucesos
- E:** Número de conatos
- F:** Número de quemas
- G:** Número de incendios
- H:** Superficie arbolada quemada
- I:** Superficie rasa quemada
- J:** Número de especialistas (técnicos, agentes forestales...)
- K:** Número de trabajadores de la brigada
- L:** Número de medios (terrestres, aéreos...)
- M:** Número de días de duración del ciclo
- N:** Valor de la medida heurística de siniestralidad del ciclo
- P:** Prototipo de evolución de la siniestralidad: E (escasamente progresiva), M (medianamente progresiva), A (altamente progresiva)

Data mining

En esta parte final del proceso de KDD, lo que se pretende es aplicar funciones de resumen sobre los ciclos de cada tipo, con el fin de obtener los tres prototipos de evolución de la siniestralidad, para poder evaluar casos reales y llegar a predecir el comportamiento y las necesidades en los siniestros de días sucesivos.

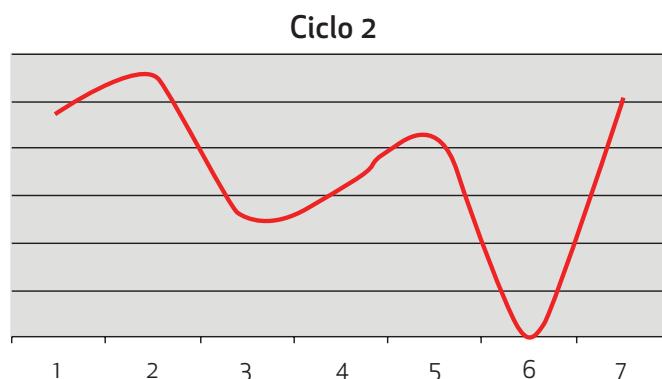
Se presenta cada ciclo con los valores relevantes de cada uno de los días y con un gráfico que representa la evolución de la siniestralidad en el tiempo (en días), tomando como dato la media de todos los valores de cada día (ocurrencia diaria), normalizados en el intervalo $[0, 10]$. El patrón de este gráfico sería el presentado al principio de este proceso.

- El **primer sector** representa un lento crecimiento de la siniestralidad en los días inmediatamente posteriores al periodo de lluvias. Durará desde el inicio del ciclo hasta que la ocurrencia diaria alcance el valor 4.
- El **segundo sector** expresa un alto crecimiento de la siniestralidad, especialmente en cuanto al número de incendios. Será desde que la ocurrencia diaria haya alcanzado el valor 4 por primera vez hasta que alcance el valor 7.
- En el **tercer sector** se quiere representar una estabilización en cuanto al número, pero un progresivo aumento en la peligrosidad de los fuegos. Los valores irán desde que la ocurrencia diaria haya alcanzado el valor 7 por primera vez hasta el final.
- Siniestralidad escasamente progresiva

Los ciclos 0, 1, 4, 5, 6, 10, 17, 18, 19, 20, 21, 23, 29, 30, 31 y 36 no se tienen en cuenta por ser de uno o dos días con incendios y, por lo tanto, ser nula su representatividad. Teniendo esto en cuenta, los ciclos que se analizan son los siguientes (en el eje X se representan los días del ciclo y en Y el valor de la ocurrencia diaria):

Tabla 9

Ciclos de siniestralidad escasamente progresiva

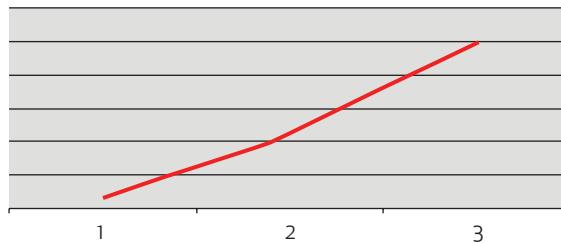


Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
24-01-1991	2	0	0,027	6	2	0	4,7	2
25-01-1991	2	0,04	0,05	2	2	0	5,6	2
26-01-1991	1	0,015	0	2	1	0	2,6	2
27-01-1991	1	0	0,01	0	1	1	3,2	2
28-01-1991	2	0,005	0,035	1	3	0	4,2	2
30-01-1991	1	0	0,15	1	1	1	5,0	2

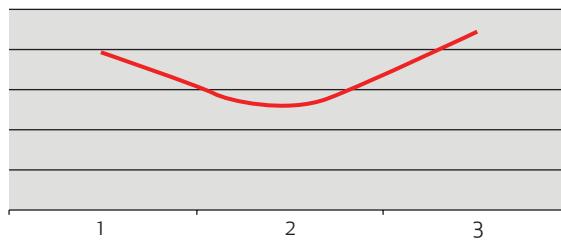
>>>

Tabla 9

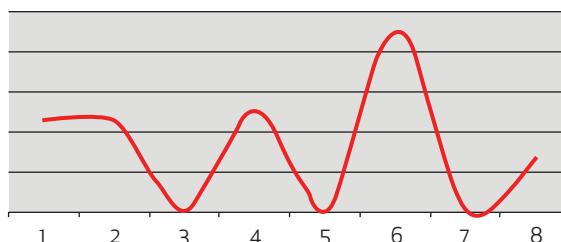
Continuación

Ciclo 3

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
23-02-1991	1	0	0,02	0	1	0	0,6	1
24-02-1991	4	0,005	0,19	4	4	0	4,4	2
25-02-1991	7	0,018	0,185	13	8	2	10,0	3

Ciclo 8

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
09-04-1991	9	0,03	0,151	10	10	4	5,5	2
10-04-1991	5	0,01	0,095	6	5	5	3,5	2
11-04-1991	6	0,128	0,16	12	8	5	5,2	2

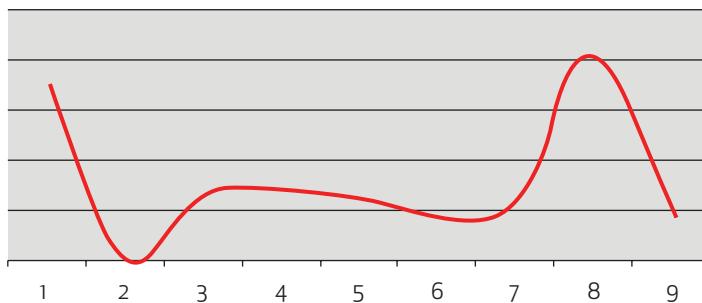
Ciclo 12

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
01-06-1991	1	0	0,03	1	1	1	4,4	2
02-06-1991	1	0	0,03	1	1	1	4,4	2
03-06-1991	0	0	0	0	0	0	0,0	2
04-06-1991	2	0,001	0,006	1	2	1	4,9	2
05-06-1991	0	0	0	0	0	0	0,0	2
06-06-1991	2	0,01	0,01	7	2	2	8,9	3
07-06-1991	0	0	0	0	0	0	0,0	3
08-06-1991	1	0,001	0	3	1	0	2,5	3

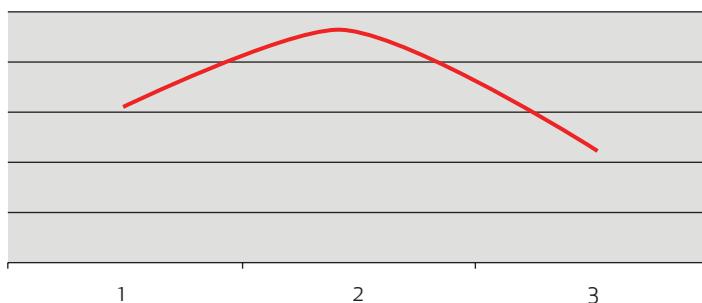
>>>

Tabla 9

Continuación

Ciclo 13

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
11-06-1991	3	0,083	0,08	6	4	5	6,9	2
12-06-1991	0	0	0	0	0	0	0,0	2
13-06-1991	1	0,02	0,03	4	2	2	2,6	2
14-06-1991	2	0,01	0,001	4	5	2	2,9	2
15-06-1991	3	0	0,009	1	3	3	2,6	2
16-06-1991	2	0,001	0,002	1	3	2	1,9	2
17-06-1991	2	0,005	0	1	3	3	2,1	2
18-06-1991	4	0,05	0,028	14	8	11	8,3	3
19-06-1991	2	0	0,006	2	2	2	1,9	3

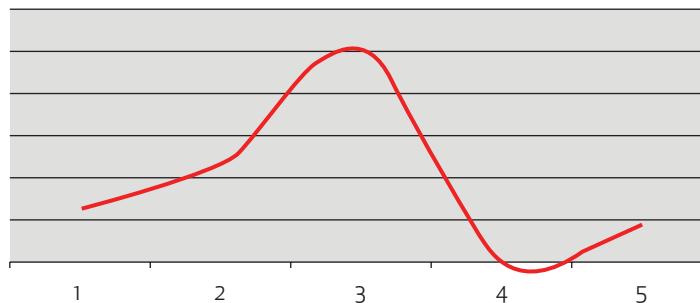
Ciclo 22

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
05-12-1991	4	0	0,21	5	9	2	6,2	2
06-12-1991	6	0,01	0,116	9	9	4	9,3	3
07-12-1991	6	0	0,07	3	7	1	4,5	3

>>>

Tabla 9

Continuación

Ciclo 25

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
16-01-1992	1	0	0,5	1	2	1	2,6	1
17-01-1992	2	0,51	0,01	2	2	1	4,8	2
18-01-1992	5	0,5	2	3	7	2	10,0	3
19-01-1992	0	0	0	0	0	0	0,0	3
20-01-1992	1	0	0,2	2	1	0	1,8	3

Analizando estos ciclos y aplicando funciones que los resuman, se llega a la definición del prototipo de **siniestralidad escasamente progresiva**:

Tabla 10

Prototipo de siniestralidad escasamente progresiva

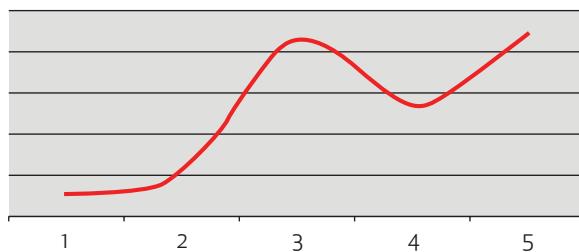
	Sector 1	Sector 2	Sector 3
Media de días	1	4	2
Media de incendios/día	1	2	4
Mínimo de incendios/día	1	1	1
Máximo de incendios/día	1	4	9
Núm. de especialistas/día	1	2	6
Núm. de trabajadores/día	2	2	5
Núm. de medios/día	1	1	3

- Siniestralidad medianamente progresiva

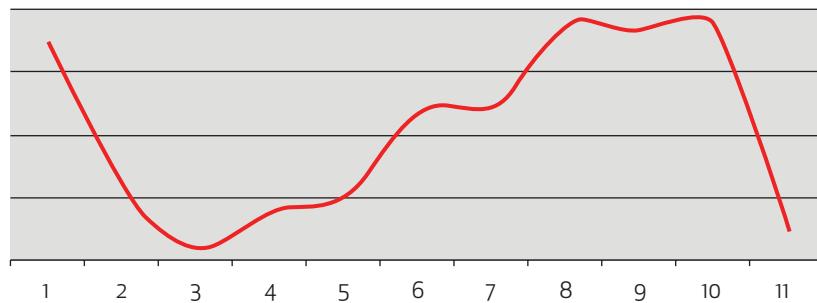
Los ciclos que se analizan son los siguientes (en el eje X se representan los días del ciclo y en Y el valor de la ocurrencia diaria):

Tabla 11

Ciclos de siniestralidad medianamente progresiva

Ciclo 7

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
29-03-1991	3	0	0,1	1	3	1	1,1	1
30-03-1991	4	0,035	0,04	2	4	3	2,1	1
31-03-1991	14	0,079	0,569	11	19	12	8,6	3
01-04-1991	11	0,011	0,23	12	12	11	5,5	3
02-04-1991	13	0,088	0,27	15	22	18	9,0	3

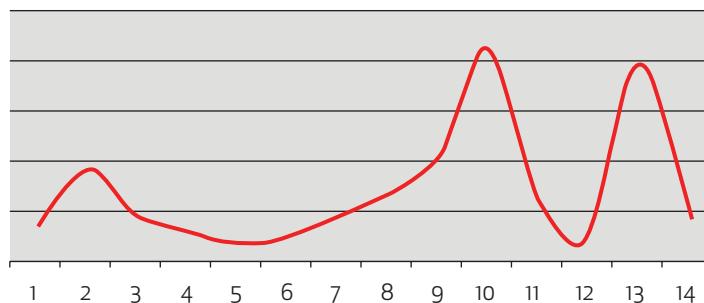
Ciclo 9

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
14-04-1991	4	0,005	0,235	18	8	8	6,9	2
15-04-1991	4	0,005	0,013	0	5	3	2,4	2
16-04-1991	1	0	0,005	0	1	0	0,4	2
17-04-1991	2	0,01	0,005	3	4	1	1,6	2
18-04-1991	2	0,012	0,013	4	4	2	2,0	2
19-04-1991	6	0,04	0,117	6	9	2	4,8	2
20-04-1991	6	0,013	0,07	10	9	4	4,9	2
21-04-1991	7	0,158	0,16	13	9	4	7,5	3
22-04-1991	7	0,074	0,036	21	10	8	7,3	3
23-04-1991	8	0,075	0,11	11	11	8	7,4	3
24-04-1991	1	0	0,06	1	1	0	0,9	3

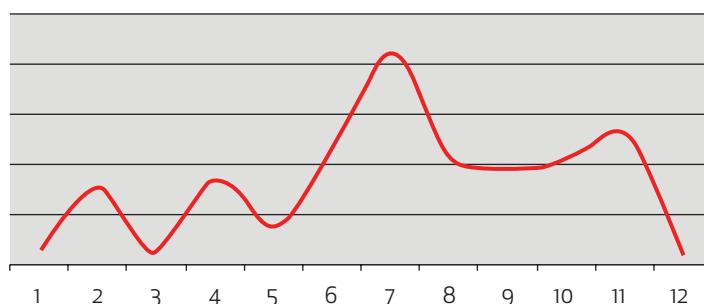
>>>

Tabla 11

Continuación

Ciclo 14

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
22-06-1991	2	0,015	0,01	4	5	2	1,5	1
23-06-1991	6	0,01	0,053	8	12	9	3,7	1
24-06-1991	4	0	0,016	4	5	5	1,9	1
25-06-1991	2	0	0,035	6	2	3	1,3	1
26-06-1991	1	0	0,01	1	3	2	0,7	1
27-06-1991	3	0,003	0,003	0	3	0	0,9	1
28-06-1991	3	0	0,124	3	5	3	1,7	1
29-06-1991	6	0,009	0,024	7	7	5	2,8	1
30-06-1991	5	0	0,154	14	11	10	4,0	2
01-jul-91	6	0,124	0,505	26	19	15	8,5	3
02-jul-91	5	0	0,051	5	5	8	2,5	3
03-jul-91	2	0	0,006	2	3	2	0,9	3
04-jul-91	9	0,017	0,813	16	17	19	7,7	3
06-jul-91	3	0	0,037	7	5	4	1,9	3

Ciclo 16

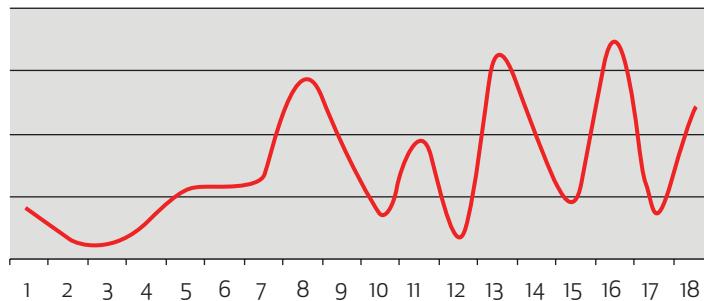
Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
14-09-1991	2	0	0,006	3	3	4	0,6	1
15-09-1991	5	0,003	0,093	13	15	20	3,1	1
16-09-1991	2	0	0,01	1	3	3	0,5	1
17-09-1991	8	0,002	0,123	8	20	23	3,5	1

>>>

Tabla 11

Continuación

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
18-09-1991	4	0,008	0,035	6	7	5	1,5	1
19-09-1991	11	0,012	0,068	18	22	22	4,3	2
20-09-1991	19	0,01	0,286	31	43	42	8,4	3
21-09-1991	6	0,06	0,12	12	14	10	4,4	3
22-09-1991	10	0,001	0,155	13	20	18	3,9	3
23-09-1991	11	0	0,102	13	25	23	4,1	3
24-09-1991	8	0,015	0,322	12	22	25	5,3	3
25-09-1991	1	0	0,015	2	2	3	0,5	3

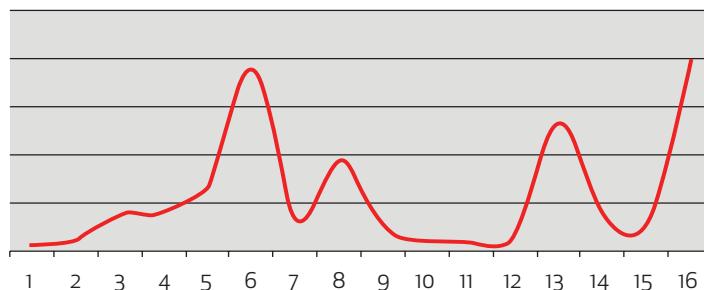
Ciclo 32

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
29-06-1992	1	0	1	1	1	1	1,7	1
01-07-1992	1	0	0,1	0	1	1	0,6	1
02-07-1992	1	0	0	0	1	0	0,4	1
03-07-1992	2	0	0,11	1	2	1	1,1	1
04-07-1992	2	0	0,51	5	2	2	2,2	1
05-07-1992	4	0	0,4	0	4	2	2,2	1
06-07-1992	3	0	0,14	7	3	3	2,7	1
07-07-1992	4	0,4	1,3	7	6	7	5,8	2
09-07-1992	1	1,2	0	4	5	3	3,7	2
10-07-1992	3	0	0,13	0	3	1	1,4	2
11-07-1992	4	0,25	0,19	6	6	4	3,8	2
12-07-1992	1	0	0,04	1	1	1	0,7	2
13-07-1992	4	0	1,22	13	7	9	6,5	2
14-07-1992	5	0,13	0,18	4	6	4	3,6	2
15-07-1992	3	0	0,34	1	3	2	1,9	2
16-07-1992	8	0,1	1,14	8	11	7	7,0	3
17-07-1992	3	0	0,13	1	3	1	1,5	3
18-07-1992	3	0	1,64	6	5	5	4,7	3

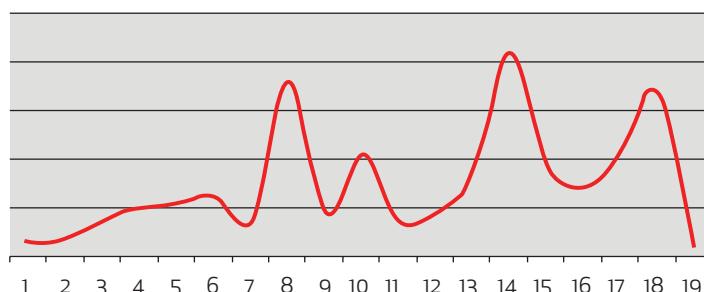
>>>

Tabla 11

Continuación

Ciclo 34

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
12-08-1992	1	0,01	0	1	1	0	0,2	1
13-08-1992	2	0	0,16	1	2	2	0,5	1
14-08-1992	6	0,01	0,64	3	8	6	1,6	1
15-08-1992	4	0	1,57	6	8	4	1,6	1
16-08-1992	8	0	1,17	9	14	11	2,7	1
17-08-1992	17	0,25	9,17	27	25	39	7,6	3
18-08-1992	4	0	0,73	4	6	7	1,4	3
19-08-1992	7	0,85	1,75	16	16	16	3,7	3
20-08-1992	2	0,03	0,65	5	3	4	0,9	3
21-08-1992	2	0	0,01	2	2	1	0,5	3
22-08-1992	2	0	0,24	0	2	1	0,4	3
23-08-1992	3	0	0,16	0	3	3	0,6	3
24-08-1992	9	0	6,91	18	26	23	5,3	3
25-08-1992	3	0	0,39	8	6	6	1,4	3
26-08-1992	3	0	0,26	7	3	3	1,1	3
27-08-1992	7	5	13,65	24	30	16	7,9	3

Ciclo 35

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
02-09-1992	1	0,1	0	3	2	1	0,7	1
04-09-1992	1	0	0,2	4	1	2	0,8	1
05-09-1992	3	0,04	0,54	1	6	4	1,5	1
06-09-1992	5	0	0,1	2	8	5	2,0	1

>>>

Tabla 11

Continuación

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
07-09-1992	3	0,28	0,12	6	9	4	2,2	1
08-09-1992	3	0,13	0,06	15	6	4	2,5	1
09-09-1992	3	0,15	0,05	3	5	4	1,5	1
10-09-1992	9	0	5,41	14	20	21	7,1	3
11-09-1992	2	0,5	1	4	4	3	1,9	3
12-09-1992	6	0	0,86	20	10	9	4,2	3
13-09-1992	3	0,02	0,15	3	4	6	1,5	3
14-09-1992	4	0	0,25	2	5	5	1,6	3
15-09-1992	4	0	2,99	7	7	6	3,0	3
16-09-1992	10	1,87	1,34	22	24	16	8,3	3
17-09-1992	4	0,7	1,25	5	10	13	3,7	3
18-09-1992	4	0,02	1,02	10	8	7	2,8	3
19-09-1992	6	0	1,11	14	12	13	4,2	3
20-09-1992	8	0,11	6,14	18	13	18	6,8	3
21-09-1992	1	0	0,15	1	1	1	0,4	3

Analizando estos ciclos y aplicando funciones que los resuman, se llega a la definición del prototipo de **siniestralidad medianamente progresiva**:

Tabla 12

Prototipo de siniestralidad medianamente progresiva

	Sector 1	Sector 2	Sector 3
Media de días	6	2	6
Media de incendios/día	3	4	6
Mínimo de incendios/día	1	1	1
Máximo de incendios/día	8	11	19
Núm. de especialistas/día	4	6	11
Núm. de trabajadores/día	5	6	12
Núm. de medios/día	4	5	11

- Siniestralidad altamente progresiva

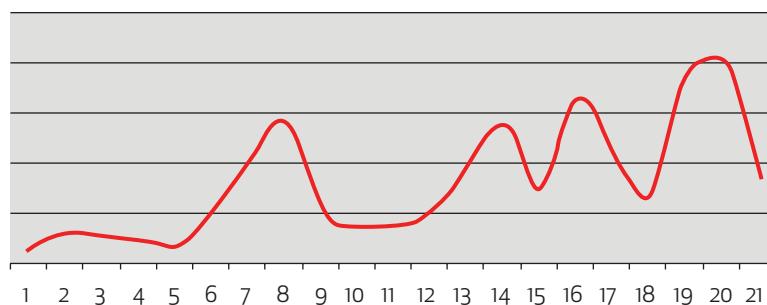
Los ciclos que se analizan son los siguientes (en el eje X se representan los días del ciclo y en Y el valor de la ocurrencia diaria):

- El **primer sector** durará desde el inicio del ciclo hasta que la ocurrencia diaria alcance el valor 4.
- El **segundo sector** será desde que la ocurrencia diaria haya alcanzado el valor 4 por primera vez hasta que alcance el valor 7.

- En el **tercer sector** los valores irán desde que la ocurrencia diaria haya alcanzado el valor 7 por primera vez hasta el final.

Tabla 13

Ciclos de siniestralidad altamente progresiva

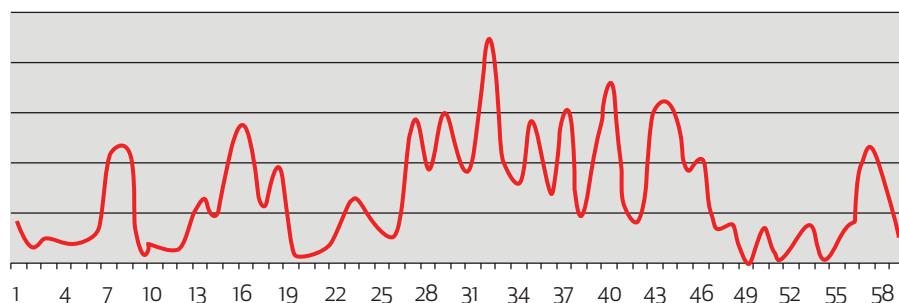
Ciclo 11

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
08-05-1991	1	0	0,001	5	2	1	0,6	1
09-05-1991	1	0,01	0,09	6	3	2	1,1	1
10-05-1991	1	0,05	0,01	5	2	1	1,1	1
12-05-1991	3	0,007	0,01	1	3	3	1,0	1
13-05-1991	2	0,002	0,007	3	2	1	0,7	1
14-05-1991	5	0,002	0,032	6	7	5	2,0	1
15-05-1991	6	0,03	0,33	13	10	10	3,9	1
16-05-1991	7	0,176	0,097	16	15	10	5,6	2
17-05-1991	4	0,002	0,265	7	6	5	2,2	2
18-05-1991	4	0,004	0,011	5	6	2	1,5	2
19-05-1991	3	0,03	0,212	4	5	1	1,7	2
20-05-1991	5	0,01	0,025	7	6	4	2,0	2
21-05-1991	8	0,03	0,28	12	11	4	3,6	2
22-05-1991	8	0,052	0,506	18	17	10	5,5	2
23-05-1991	5	0,044	0,025	19	7	4	3,0	2
24-05-1991	7	0,152	0,275	33	16	10	6,6	2
25-05-1991	8	0,106	0,078	17	10	9	4,6	2
26-05-1991	7	0,013	0,039	7	8	6	2,6	2
27-05-1991	9	0,096	1,063	23	14	15	7,4	3
28-05-1991	13	0,16	0,155	25	21	18	8,0	3
29-05-1991	6	0,014	0,13	16	14	5	3,5	3

>>>

Tabla 13

Continuación

Ciclo 15

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
10-07-1991	8	0,001	0,039	14	8	7	1,4	1
12-07-1991	3	0,021	0,081	7	4	5	0,7	1
13-07-1991	4	0,022	0,04	9	6	9	1,0	1
14-07-1991	2	0,05	0,005	6	7	7	0,8	1
15-07-1991	3	0	0,03	5	6	9	0,8	1
16-07-1991	4	0,004	0,04	10	11	9	1,2	1
17-07-1991	17	0,004	0,188	32	33	35	4,2	2
18-07-1991	15	0,094	0,495	38	37	35	4,6	2
19-07-1991	2	0	0,004	2	2	3	0,3	2
20-07-1991	2	0,02	0	4	7	6	0,7	2
21-07-1991	3	0	0,046	3	5	4	0,6	2
22-07-1991	5	0,008	0,023	5	7	4	0,8	2
23-07-1991	6	0,011	0,283	26	18	20	2,5	2
24-07-1991	7	0,08	0,022	13	15	14	1,8	2
25-07-1991	17	0,01	0,274	33	41	45	4,8	2
26-07-1991	22	0,015	0,16	43	44	43	5,5	2
27-07-1991	9	0,027	0,121	18	20	18	2,4	2
28-07-1991	12	0,02	0,243	40	28	23	3,7	2
30-07-1991	1	0,022	0,015	6	4	2	0,5	2
01-08-1991	2	0	0,013	2	3	1	0,3	2
02-08-1991	2	0	0,006	3	4	4	0,5	2
03-08-1991	4	0	0,036	14	10	10	1,3	2
04-08-1991	7	0,135	0,098	20	19	23	2,5	2
05-08-1991	4	0,025	0,222	12	12	20	1,7	2
06-08-1991	5	0	0,034	4	12	13	1,2	2

>>>

Tabla 13

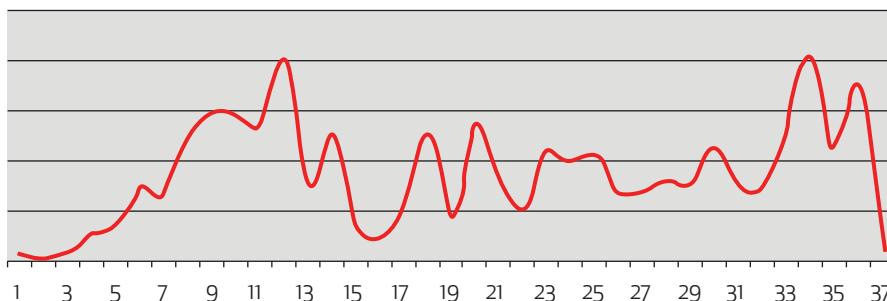
Continuación

Día	N.º inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
07-08-1991	7	0,003	0,01	8	11	9	1,3	2
08-08-1991	17	0,065	0,684	46	49	41	5,6	2
09-08-1991	13	0,061	0,088	32	32	27	3,7	2
10-08-1991	12	1,622	1,692	27	35	33	6,1	2
11-08-1991	12	0,083	0,249	32	30	29	3,7	2
12-08-1991	14	0,021	0,273	38	30	27	3,9	2
13-08-1991	20	1,101	3,505	53	61	52	9,1	3
14-08-1991	13	0,161	0,459	31	29	29	3,9	3
15-08-1991	11	0,013	0,201	24	26	28	3,2	3
16-08-1991	21	0,109	0,364	34	51	46	5,6	3
17-08-1991	12	0,005	0,121	13	26	27	2,8	3
18-08-1991	21	0,048	0,248	48	60	45	6,1	3
19-08-1991	11	0	0,079	4	21	17	2,0	3
20-08-1991	19	0,021	0,11	25	40	36	4,3	3
21-08-1991	21	0,456	1,83	42	65	41	7,1	3
22-08-1991	7	0,02	0,108	12	17	17	1,9	3
23-08-1991	9	0	0,061	6	19	15	1,8	3
24-08-1991	24	0,077	0,877	38	50	51	6,2	3
25-08-1991	22	0,017	0,463	33	49	53	5,7	3
26-08-1991	16	0,054	0,099	27	31	26	3,7	3
27-08-1991	18	0,004	0,194	31	41	21	4,0	3
28-08-1991	6	0	0,142	12	13	10	1,5	3
29-08-1991	4	0,01	0,02	9	15	15	1,4	3
30-08-1991	1	0	0,001	0	2	0	0,1	3
01-sep-1991	6	0,005	0,064	5	16	13	1,4	3
02-sep-1991	1	0,002	0	2	2	1	0,2	3
03-sep-1991	2	0	0,004	5	5	3	0,5	3
04-sep-1991	5	0	0,065	6	14	19	1,5	3
05-sep-1991	1	0	0,01	0	1	0	0,1	3
06-sep-1991	3	0,01	0,004	9	8	6	0,9	3
07-sep-1991	7	0,01	0,03	12	13	11	1,6	3
08-sep-1991	16	0,01	0,171	26	43	42	4,4	3
09-sep-1991	11	0,033	0,117	22	26	29	3,1	3
10-sep-1991	5	0	0,061	6	12	8	1,1	3

>>>

Tabla 13

Continuación

Ciclo 24

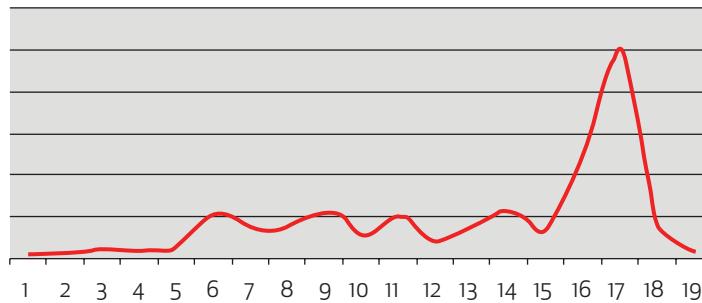
Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
15-02-1992	1	0	1,5	2	2	1	0,2	1
16-02-1992	1	0,2	0	1	2	1	0,2	1
18-02-1992	3	0	2,7	1	4	2	0,3	1
19-02-1992	9	0,1	10,9	4	13	5	1,1	1
20-02-1992	13	0,3	13,13	6	16	5	1,4	1
21-02-1992	19	6,1	27,52	11	29	9	2,9	1
22-02-1992	22	5,9	15,61	12	30	7	2,7	1
23-02-1992	42	2,5	66,7	14	61	18	5,1	2
24-02-1992	43	7,1	78,63	21	62	15	5,9	2
25-02-1992	35	3,5	82,31	30	54	17	5,7	2
26-02-1992	43	4,2	38,11	36	61	18	5,5	2
27-02-1992	60	21,9	39,9	29	84	27	7,9	3
28-02-1992	29	3,8	22,31	9	41	7	3,0	3
29-02-1992	34	7,7	66,1	13	57	15	5,1	3
01-03-1992	9	0,1	19,8	5	12	3	1,2	3
02-03-1992	6	1	7	3	13	4	0,9	3
03-03-1992	15	5,9	8,6	11	25	8	2,3	3
04-03-1992	36	9,1	38,1	20	60	17	5,1	3
05-03-1992	17	0,6	16	9	22	7	1,9	3
06-03-1992	35	1,9	56,3	36	47	26	5,5	3
07-03-1992	20	9,7	15,9	11	26	11	3,0	3
08-03-1992	21	2,7	8,7	4	26	10	2,0	3
09-03-1992	31	6,2	40,8	17	51	17	4,5	3
10-03-1992	24	4,7	50,8	18	36	17	4,1	3
11-03-1992	24	15	25,2	12	43	17	4,3	3
12-03-1992	15	6,7	12,1	14	30	14	2,8	3
13-03-1992	25	2,9	23,2	9	41	14	3,1	3
14-03-1992	20	2,6	25,45	19	38	13	3,2	3

>>>

Tabla 13

Continuación

15-03-1992	24	3,8	9,2	10	37	21	3,1	3
16-03-1992	23	8,6	23,75	24	50	24	4,6	3
17-03-1992	22	2,35	19,1	16	32	13	2,9	3
18-03-1992	23	4,6	14,5	13	34	14	3,0	3
19-03-1992	41	4,9	56,7	27	56	23	5,7	3
20-03-1992	44	15,95	43	45	78	43	8,2	3
21-03-1992	36	6,7	28,4	18	54	21	4,6	3
22-03-1992	47	3,8	61	39	72	30	6,9	3
23-03-1992	4	0	5,5	1	7	1	0,4	3

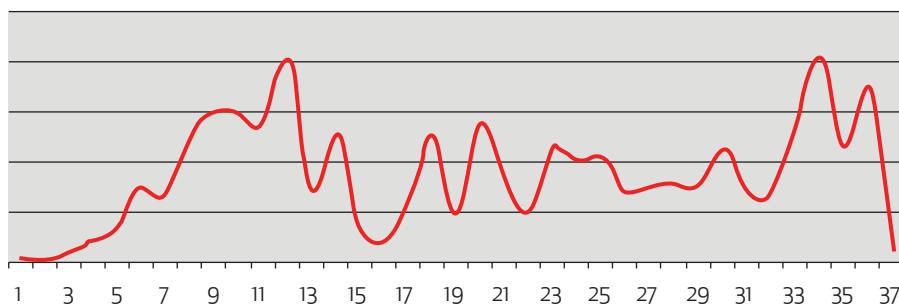
Ciclo 26

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
23-01-1992	5	0	3,6	4	6	0	0,3	1
24-01-1992	3	0	4,6	4	5	2	0,3	1
25-01-1992	6	1,1	2,5	7	12	3	0,5	1
26-01-1992	8	0,5	8,8	3	13	4	0,5	1
27-01-1992	6	0,3	19	4	17	3	0,6	1
28-01-1992	28	10,1	52,5	28	39	8	2,2	1
29-01-1992	22	5	33,2	12	32	7	1,5	1
30-01-1992	23	2	26,45	18	33	6	1,5	1
31-01-1992	31	11,7	50,53	21	43	12	2,3	1
01-02-1992	16	5,9	12,4	7	17	5	0,9	1
02-02-1992	27	7,4	25,65	28	40	8	2,0	1
03-02-1992	11	1	34,63	8	15	4	0,8	1
04-02-1992	27	1,4	19,6	18	38	4	1,6	1
05-02-1992	27	1,5	125,8	27	45	8	2,4	1
06-02-1992	30	3	24,21	4	38	9	1,6	1
07-02-1992	53	25,2	128,61	36	87	27	4,5	2
08-02-1992	85	115,1	425,56	88	124	61	10,0	3
09-02-1992	13	14,8	66,5	11	25	11	1,6	3
11-02-1992	3	0,3	0,6	8	6	5	0,4	3

>>>

Tabla 13

Continuación

Ciclo 27

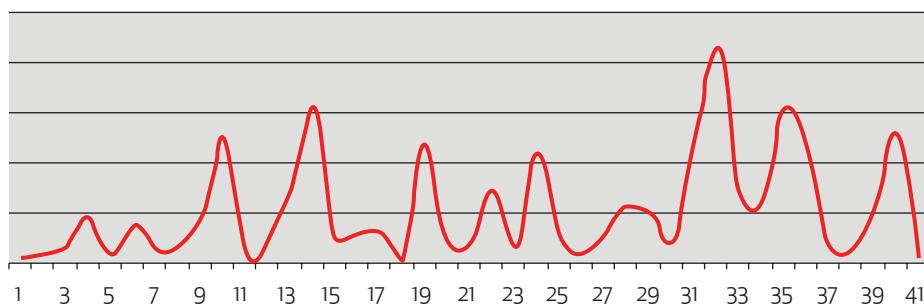
Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
15-02-1992	1	0	1,5	2	2	1	0,2	1
16-02-1992	1	0,2	0	1	2	1	0,2	1
18-02-1992	3	0	2,7	1	4	2	0,3	1
19-02-1992	9	0,1	10,9	4	13	5	1,1	1
20-02-1992	13	0,3	13,13	6	16	5	1,4	1
21-02-1992	19	6,1	27,52	11	29	9	2,9	1
22-02-1992	22	5,9	15,61	12	30	7	2,7	1
23-02-1992	42	2,5	66,7	14	61	18	5,1	2
24-02-1992	43	7,1	78,63	21	62	15	5,9	2
25-02-1992	35	3,5	82,31	30	54	17	5,7	2
26-02-1992	43	4,2	38,11	36	61	18	5,5	2
27-02-1992	60	21,9	39,9	29	84	27	7,9	3
28-02-1992	29	3,8	22,31	9	41	7	3,0	3
29-02-1992	34	7,7	66,1	13	57	15	5,1	3
01-03-1992	9	0,1	19,8	5	12	3	1,2	3
02-03-1992	6	1	7	3	13	4	0,9	3
03-03-1992	15	5,9	8,6	11	25	8	2,3	3
04-03-1992	36	9,1	38,1	20	60	17	5,1	3
05-03-1992	17	0,6	16	9	22	7	1,9	3
06-03-1992	35	1,9	56,3	36	47	26	5,5	3
07-03-1992	20	9,7	15,9	11	26	11	3,0	3
08-03-1992	21	2,7	8,7	4	26	10	2,0	3
09-03-1992	31	6,2	40,8	17	51	17	4,5	3
10-03-1992	24	4,7	50,8	18	36	17	4,1	3
11-03-1992	24	15	25,2	12	43	17	4,3	3
12-03-1992	15	6,7	12,1	14	30	14	2,8	3

>>>

Tabla 13

Continuación

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
13-03-1992	25	2,9	23,2	9	41	14	3,1	3
14-03-1992	20	2,6	25,45	19	38	13	3,2	3
15-03-1992	24	3,8	9,2	10	37	21	3,1	3
16-03-1992	23	8,6	23,75	24	50	24	4,6	3
17-03-1992	22	2,35	19,1	16	32	13	2,9	3
18-03-1992	23	4,6	14,5	13	34	14	3,0	3
19-03-1992	41	4,9	56,7	27	56	23	5,7	3
20-03-1992	44	15,95	43	45	78	43	8,2	3
21-03-1992	36	6,7	28,4	18	54	21	4,6	3
22-03-1992	47	3,8	61	39	72	30	6,9	3
23-03-1992	4	0	5,5	1	7	1	0,4	3

Ciclo 28

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
10-04-1992	1	0,4	0,5	1	2	1	0,3	1
11-04-1992	1	0	1,5	1	2	1	0,3	1
12-04-1992	3	0	4,1	2	5	1	0,7	1
14-04-1992	6	3,3	9	9	11	3	1,9	1
15-04-1992	1	0	0,5	0	1	0	0,1	1
16-04-1992	6	2,2	4,8	8	10	3	1,5	1
17-04-1992	3	0	1,2	4	4	1	0,6	1
18-04-1992	3	0,5	0,5	4	4	1	0,6	1
19-04-1992	8	0,7	7,4	2	9	6	1,6	1
20-04-1992	14	1,3	37,6	21	24	13	5,0	2
21-04-1992	1	0,1	0,2	3	1	1	0,3	2

>>>

Tabla 13

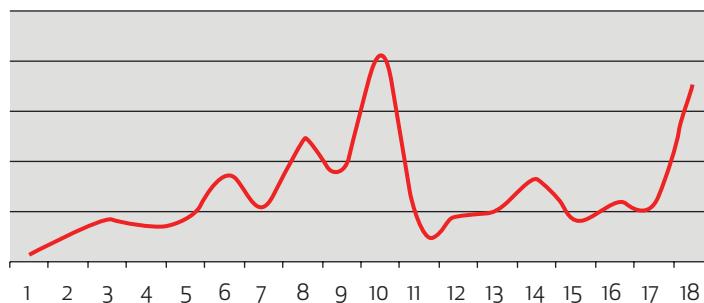
Continuación

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
22-04-1992	5	0,2	2,5	7	6	4	1,1	2
23-04-1992	9	0,6	6,8	17	21	9	2,8	2
24-04-1992	16	19,1	29,5	20	34	18	6,3	2
25-04-1992	5	0,1	0,6	4	5	4	0,9	2
26-04-1992	5	0,5	2,7	9	7	5	1,3	2
27-04-1992	3	0,1	8,5	6	5	4	1,2	2
28-04-1992	1	0	2	0	3	1	0,3	2
29-04-1992	5	30	28,6	10	14	10	4,6	2
30-04-1992	5	0,1	1,2	3	7	2	0,9	2
02-05-1992	2	2	0,1	7	4	2	0,7	2
03-05-1992	6	10,1	12,2	13	14	8	2,9	2
04-05-1992	3	2	0,7	2	4	1	0,6	2
05-05-1992	11	2,7	20,8	23	24	17	4,3	2
06-05-1992	4	1	3,6	2	7	2	0,9	2
07-05-1992	2	0	0,3	3	2	2	0,4	2
08-05-1992	5	0,1	2,7	6	8	5	1,2	2
09-05-1992	7	1,3	6	14	13	9	2,2	2
10-05-1992	7	0,3	11,6	5	13	9	2,1	2
11-05-1992	3	1,6	0,7	3	3	5	0,8	2
12-05-1992	15	2,2	7,9	17	29	20	4,2	2
13-05-1992	23	22,1	12,9	50	46	34	8,5	3
14-05-1992	5	4,1	9,9	14	16	9	2,5	3
15-05-1992	8	1,2	1	8	19	14	2,3	3
16-05-1992	17	2,2	32,9	19	34	27	6,0	3
17-05-1992	22	2	6,6	17	35	22	4,9	3
18-05-1992	1	0	2,5	3	2	3	0,5	3
19-05-1992	2	0	0,5	3	3	2	0,5	3
20-05-1992	6	0,5	6,4	7	10	8	1,7	3
21-05-1992	17	24,5	4,4	14	31	16	5,2	3
23-05-1992	1	0,1	0	3	2	2	0,3	3

>>>

Tabla 13

Continuación

Ciclo 33

Día	Núm. inc.	Sup. arb.	Sup. rasa	Espec.	Trabaj.	Medios	Ocurr. diaria	Sector
21-07-1992	1	0	0,1	2	2	1	0,3	1
22-07-1992	3	0,45	0,25	6	6	5	1,0	1
23-07-1992	3	0,8	1,8	8	8	11	1,7	1
24-07-1992	5	0	1,22	6	7	9	1,5	1
25-07-1992	5	0,4	1,18	12	6	8	1,7	1
26-07-1992	8	1,15	3,62	16	19	21	3,5	1
27-07-1992	2	0	13,8	7	8	8	2,2	1
28-07-1992	9	1,95	23,39	12	20	20	4,9	2
29-07-1992	10	2,35	6,71	17	18	13	3,6	2
30-07-1992	23	15,4	4,47	37	44	23	8,2	3
31-07-1992	4	0,75	0,08	3	6	7	1,1	3
01-08-1992	8	1,25	1,15	4	12	8	1,8	3
02-08-1992	7	0,5	0,72	13	11	7	2,0	3
03-08-1992	9	0,3	5,21	21	16	13	3,3	3
04-08-1992	5	0,16	2,62	4	10	9	1,6	3
05-08-1992	6	0,27	3,41	7	20	10	2,3	3
06-08-1992	6	0	5,39	9	11	12	2,3	3
07-08-1992	19	5,07	7,78	29	42	32	7,0	3

Analizando estos ciclos y aplicando funciones que los resuman, se llega a la definición del prototipo de **siniestralidad altamente progresiva**:

Tabla 14
Prototipo de siniestralidad altamente progresiva

	Sector 1	Sector 2	Sector 3
Media de días	8	13	15
Media de incendios/día	9	12	19
Mínimo de incendios/día	1	1	1
Máximo de incendios/día	31	53	85
Núm. de especialistas/día	8	16	17
Núm. de trabajadores/día	13	21	33
Núm. de medios/día	5	13	18

Estas definiciones de los 3 prototipos pueden ser útiles para establecer modelos de prevención de incendios forestales a partir de la predicción de cómo podría evolucionar su posibilidad de ocurrencia en las fechas siguientes a una dada.

Descarga de archivo

También se puede desarrollar un sistema sofisticado de organización de recursos con reajuste diario, como se puede ver en detalle en la tesis doctoral que puede descargarse en *Campus Virtual > Aula de la asignatura > Recursos y materiales* (archivo: [Contribución al estudio experimental de la predicción basada en categorías deformables](#)).

4.2. Computación suave y lógica borrosa para la búsqueda y recuperación de información (en la Web)

Los **sistemas de recuperación de información** son un tema actualmente muy destacado en las ciencias de la computación y la inteligencia artificial.

Descarga de archivo

Se pueden consultar más detalles en el libro *Búsqueda eficaz de información en la Web*, disponible en *Campus Virtual > Aula de la asignatura > Recursos y materiales* (archivo: [Búsqueda eficaz de información en la Web](#)).

Habitualmente, un sistema de recuperación de información es definido como el proceso que trata la representación, el almacenamiento, la organización y el acceso de elementos de información. Es decir, es un **sistema capaz de almacenar, recuperar y mantener información**.

Pero podríamos plantearnos qué representa el concepto de **información** en este contexto. Se entiende por información cualquier elemento susceptible de ser recuperado, lo que incluye principalmente texto (incluidos números y fechas), imágenes, audio, vídeo y otros objetos multimedia.

Sin embargo, el tipo principal de objeto recuperable hasta el momento siempre ha sido el **texto**, motivado especialmente por su facilidad de manipulación en comparación con los objetos multimedia, especialmente en lo que se refiere a capacidad de cómputo. Actualmente están surgiendo muchos sistemas que tratan de gestionar este tipo de objetos (diversos buscadores comerciales incluyen buscadores de imágenes), aunque de momento simplemente buscan en el texto de las etiquetas de dichos objetos multimedia, sin escudriñar realmente su contenido interno, lo que suele dar frecuentemente origen a engaños o falsos etiquetados.

En los sistemas de recuperación de información no se suele trabajar directamente con los documentos de texto sino con **representaciones más estructuradas** de estos. La forma de representar los documentos determina en gran medida las características del resto de elementos del sistema. Los modelos de representación de documentos clásicos se basan generalmente en el modelo **booleano** o en el modelo **vectorial**. En el primero, cada documento es representado por un vector donde cada posición se corresponde con cada uno de los términos susceptibles de aparecer en el documento; el valor de cada posición será 0 o 1 según si ese término aparece o no en ese documento. La esencia del modelo **vectorial** es similar, salvo que el contenido de cada componente representa algún valor que tiene que ver con la frecuencia de aparición de ese término en el documento. Claramente, estos modelos de representación de documentos son adecuados para documentos de texto, que pueden ser, por ejemplo, páginas web u otros objetos (como elementos multimedia) que estén descritos de forma textual.

Quizá el concepto más importante en recuperación de información es el de **relevancia**. Tiene que ver con cómo medir la satisfacción de un usuario con los resultados devueltos por el sistema ante una determinada pregunta (**query**). Esta medida es claramente subjetiva, ya que ante una misma **query** y el mismo resultado (documento o lista ordenada de documentos), la relevancia puede ser totalmente distinta para dos usuarios diferentes, e imposible de medir de forma precisa. Esta es una de las razones por las que cada vez se tiene más en cuenta el papel del usuario en los sistemas de recuperación de información: si se conocen los intereses de los usuarios, el sistema puede guiar la búsqueda de información hacia ellos.



Ejemplo

Supongamos que dos usuarios diferentes (U1 y U2) introducen la consulta "monitor barato" en un buscador web comercial. Si U1 habitualmente hace búsquedas en páginas de gimnasios y deportes y U2 lo hace en páginas de productos informáticos, lo más probable es que U1 esté buscando entrenadores baratos y U2 pantallas de ordenador baratas. Si se hubiesen almacenado de alguna forma estos perfiles de usuario, estas dos búsquedas podrían haber sido guiadas de formas totalmente diferentes y la relevancia de los resultados para cada usuario habría aumentado significativamente. Además, este ejemplo pone de manifiesto uno de los principales problemas de la recuperación de información, que es la propia complejidad del lenguaje natural. La palabra *monitor* es polisémica, lo que dificulta enormemente la tarea de recuperación de información cuando se usa.

Recuperar información no es recuperar datos. Cuando accedemos a una base de datos, por ejemplo, la de una biblioteca, usamos un lenguaje muy estructurado y con una semántica muy precisa. Si buscamos libros de Lope de Vega, lo pondremos en el campo “autor” de la ficha que nos proporcione el sistema, y nunca tendrá en cuenta la acepción de vega que tiene que ver con un terreno. El sistema tratará de recuperar aquellos libros de su base de datos cuyo autor es Lope de Vega. Pero, si entre sus más de mil obras, queremos localizar aquellas que contengan la palabra *rimas* en el título, el sistema puede transformar nuestra pregunta en una sentencia precisa de un lenguaje que entienda la base de datos (por ejemplo, SQL, de *structured query language* en inglés) y que represente algo como “Busca las obras de nuestra base de datos cuyo autor es Lope de Vega y cuyo título contenga la palabra *rimas*”. Con esta especificación, el sistema podría recuperar, por ejemplo, las obras *La Circe con otras rimas y prosas*, de 1624, y *Rimas, poesías* de 1604. Por el contrario, en un sistema de recuperación de información, el objeto recuperado no tiene por qué adaptarse de forma exacta a las peticiones de búsqueda. La razón fundamental es que la información que gestiona un sistema de recuperación de información está en lenguaje natural, sin estructurar, por lo que puede ser semánticamente ambigua.

Etapas del proceso de recuperación de información

Para un usuario, el proceso de recuperación de información consiste en realizar una pregunta al sistema y obtener como respuesta un conjunto de documentos ordenados. Pero, en todo sistema de recuperación de información, es necesaria la realización de una serie de pasos previos y diferenciados para poder llegar a sus respuestas. Los pasos más relevantes son los siguientes:

- 1. Indexación.** El sistema de recuperación de información crea un índice que contiene los términos que el sistema considera importantes (después de un preprocesado de cada documento) y su ubicación en los documentos.
- 2. Consulta.** El usuario formula una pregunta al sistema, en un lenguaje (formalismo) procesable por este.
- 3. Evaluación.** El sistema devuelve los resultados (documentos que satisfacen en cierto grado la demanda de información del usuario), ordenados según su (posible) relevancia con respecto a la consulta formulada.
- 4. Retroalimentación del usuario** (opcional). El sistema aprende de las diferentes consultas de un usuario, focalizando la recuperación según este conocimiento adquirido.

Estos procesos suelen ser estándar en todos los sistemas de recuperación de información. Además, esta clasificación no es cerrada, sino que, dependiendo del sistema de recuperación, pueden ser ejecutados otros nuevos métodos diferentes. Por ejemplo, los sistemas que utilizan estructuras de conocimiento adicionales a los índices de términos clásicos suelen tener procesos adicionales encargados de construir, mantener o actualizar dichas estructuras. Los sistemas de recuperación de información que se basan en el uso de perfiles de usuario pueden realizar una nueva etapa para la construcción y actualización del perfil de usuario (almacenable los términos que representan las preferencias de los usuarios con la finalidad de mejorar el comportamiento del sistema en futuras consultas).

Hay varias propuestas formales para definir un modelo de sistema de recuperación de información, pero una de las más utilizadas es la de Baeza-Yates (Baeza-Yates y Ribeiro-Neto, 1999), que define un sistema de este tipo como una cuádrupla $[D, Q, F, R(q_i, dj)]$, donde:

- D es un conjunto de vistas lógicas (o representaciones) de los documentos que forman la colección.
- Q es un conjunto compuesto por vistas lógicas (o representaciones) de las necesidades de información de los usuarios. Estas vistas se denominan consultas (queries en inglés).
- F es una forma de modelar la representación de los documentos, consultas y sus relaciones.
- R(q_i, dj) (ranking en inglés) es una función de evaluación que asigna un número real al par formado por una consulta $q_i \in Q$ y la representación de un documento $d_j \in D$. Este valor determinará el orden de aparición de los documentos de una consulta q_i .

Todos estos elementos se verán reflejados en las etapas del proceso de recuperación de información, que se detallan a continuación.

4.2.1. Indexación

Para conseguir D (conjunto de vistas lógicas o representaciones de los documentos que forman la colección), es necesaria la construcción de una **base documental** que contenga la información de los objetos que el sistema es capaz de escudriñar para poder llevar a cabo el proceso de recuperación. Si un objeto no está en esta base de datos, no podrá ser recuperado. Sin embargo, lo que maneja realmente el sistema no son los propios documentos susceptibles de ser recuperados sino una **representación** de estos. Los documentos se representan con algún formalismo, habitualmente creando un índice con el conjunto de términos significativos que aparecen en los documentos, que suele ser un subconjunto de todos los términos de los documentos.

También es necesario que el sistema de recuperación de información disponga de mecanismos que permitan **introducir un nuevo objeto** en la base documental, o bien utilizar mecanismos automáticos que se encargan de explorar el espacio de búsqueda (en nuestro caso, la Web) y añadir al índice la información sobre los términos que aparecen en estos nuevos documentos. Esto es necesario entre otras cosas porque los mecanismos de ordenación por relevancia para el usuario (se explicarán con detalle más adelante) suelen seleccionar como documentos más relevantes, entre otros criterios, aquellos que posean con más frecuencia o en determinada posición los términos que están en la consulta del usuario.

La opción más simple sería almacenar los documentos y buscar en cada uno de ellos la existencia o no de los términos, pero este planteamiento es inviable computacionalmente, debido a la enorme cantidad de información que sería necesario manejar, lo que imposibilitaría que el sistema fuese eficiente. Por tanto, es necesaria la utilización de **estructuras que almacenen información** sobre los documentos y que permitan realizar las **búsquedas** en tiempos razonablemente cortos.

Estas estructuras son lo que hoy en día se denominan índices. De todos modos, habitualmente los documentos suelen ser preprocesados antes de ser indexados para reducir el número de elementos (términos, signos de puntuación, ubicación de los términos...) a tener en cuenta y, por tanto, mejorar la eficiencia del proceso de recuperación.

Está claro que esta reducción de elementos supone una pérdida de información, que puede ser muy importante a la hora de buscar los documentos más relevantes ante una consulta, por lo que el secreto del éxito en esta etapa radica en encontrar el equilibrio justo entre la eliminación de elementos a indexar de los documentos y la eficiencia de los procesos de búsqueda en este índice.

4.2.2. Preprocesado de documentos

Las tareas que se utilizan habitualmente en el preprocesado de documentos para su indexación son las siguientes:

- Eliminación de signos de puntuación. Se eliminan los acentos, comas, puntos y demás signos de puntuación con el fin de tratar los términos de forma uniforme. Este proceso tiene el inconveniente de que se pierde esta información y no se podrán utilizar signos de puntuación en las consultas de los usuarios; además, los signos de puntuación poseen información semántica importante.



Ejemplo

Documento original: "Vivía cerca de León. Furioso con lo que le rodeaba, los hombres de la comarca lo odiaban".

Documento sin signos de puntuación: "vivía cerca de león furioso con lo que le rodeaba los hombres de la comarca lo odiaban".

- Eliminación de palabras prohibidas (stop words en inglés). En todos los idiomas hay un conjunto de palabras muy frecuentes que se usan por cuestiones lingüísticas de concordancia sintáctica entre palabras y frases. Si se considera que estas palabras no aportan ningún significado a un documento (cosa que no es cierta), sino que solo se utilizan para seguir las reglas del idioma, podrían ser eliminadas. Por ejemplo, en español podrían eliminarse artículos, preposiciones, conjunciones, algunos adverbios, etc.

Existen **listas para los diferentes idiomas** (denominadas *stop lists* en inglés) con estas palabras, que sirven como referencia para no tener en cuenta estas palabras cuando aparezcan en los documentos a la hora de ser indexados. También es frecuente la construcción dinámica de estas listas cuando se desarrollan sistemas de recuperación de información. En el ejemplo se puede ver claramente cómo cambia la semántica del texto tras eliminar estas palabras:



Ejemplo

Documento original: "Vivía cerca de León. Furioso con lo que le rodeaba, los hombres de la comarca lo odiaban".

Documento sin signos de puntuación y sin palabras prohibidas: "vivía león furioso rodeaba hombres comarca odiaban".

- Lematización (stemming en inglés). Consiste en obtener la raíz léxica (stem en inglés) de una palabra. Normalmente se ignoran las diferentes variaciones morfológicas que puede tener a la hora de indexar. En la mayoría de los casos, la raíz es una palabra sin significado, como en el caso de vivía y vivencia, cuya raíz común sería viv. Este mecanismo se suele aplicar para eliminar el sufijo de una palabra, pero no se aplica al prefijo. Esto se debe a la idea de que la raíz contiene la fuerza semántica de la palabra, y que los sufijos introducen ligeras modificaciones del concepto o tienen meramente funciones sintácticas.

El objetivo inicial de este proceso de lematización fue mejorar el rendimiento de los sistemas de recuperación de información al **reducir el número de palabras** que un sistema tenía que almacenar en el índice. Otra de las características de la lematización es que frecuentemente **favorece la exhaustividad** (*recall* en inglés) en la búsqueda. Es decir, se recuperan más palabras relacionadas léxicamente al tener la misma raíz y, por tanto, se obtiene un conjunto más elevado de términos, lo que evita perder términos potencialmente relevantes. Pero esto es a costa de una **reducción en la precisión**, porque los lenguajes naturales no suelen ser regulares en sus construcciones y, además, hay muchas palabras con la misma raíz cuyo significado no tiene nada que ver.

Por ejemplo, podría suceder que se indexen bajo la raíz *cas* los términos *casa*, *casero* y *casual*. Este fenómeno se denomina **sobrelematización**. También es posible que el mecanismo de lematización falle y obtenga raíces distintas para dos palabras semánticamente similares. Esta situación se denomina **bajolematización**. Este caso es muy frecuente en los verbos irregulares, y se podría dar, por ejemplo, con dos variaciones del verbo *haber*, para las palabras *habido* y *hayamos*, indexándose bajo raíces distintas (*habi* y *hay*).

Otro problema es que este método es dependiente del idioma y, por lo tanto, a la hora de indexar, sería necesario utilizar un **mecanismo específico para cada idioma**. Esta situación lleva asociada la utilización de una técnica para determinar el idioma. Además, este tipo de métodos funcionan bien con idiomas que tengan una sintaxis no excesivamente complicada, como el inglés, pero en cambio fallan mucho más con otro tipo de idiomas con estructuras más complejas, como el español. Por tanto, la lematización difiere mucho entre los distintos idiomas.

Se utilizan muchas técnicas en este tipo de métodos, entre las cuales destaca la utilización de **reglas y diccionarios**. Existen multitud de propuestas de lematización basadas en reglas, la mayoría de ellas para el inglés, de los cuales el clásico es el más sencillo, el **lematizador S**, que simplemente quita las terminaciones plurales. El método de lematización más famoso es el que se ha implementado en el **algoritmo de Porter**, desarrollado en la década de los ochenta, que elimina cerca de sesenta terminaciones en cinco etapas, en cada una de las cuales se elimina un tipo concreto. Para eliminar los errores más frecuentes descritos, se han desarrollado mecanismos basados en diccionarios, como **KSTEM**. Hay mucha polémica sobre la efectividad de la lematización y algunos autores afirman que esta técnica mejora la precisión y exhaustividad de las búsquedas en la medida en que las consultas (y también los documentos) sean más cortas.

Finalmente, se puede hablar del uso de los **n-gramas** (subsecuencia de n elementos de una secuencia dada). Al trabajar con n -gramas, se ignora el aspecto semántico de las palabras. La hipótesis de los mecanismos basados en n -gramas es que dos palabras relacionadas semánticamente suelen contener los mismos caracteres.



Ejemplo

Documento original: "Vivía cerca de León. Furioso con lo que le rodeaba, los hombres de la comarca lo odiaban".

Documento sin signos de puntuación, sin palabras prohibidas y tras un proceso de lematización: "viv león furios rodea hombr comarc odiaba".

- Eliminación de documentos duplicados. Muchos contenidos de páginas web están multiplicados (como mínimo duplicados) en diferentes sitios. La eliminación de estos documentos multiplicados permite mejorar el rendimiento de los programas encargados de la indexación y reducir el espacio de almacenamiento que ocupan los índices generados.

Sin embargo, la tarea de **identificar documentos iguales** o similares no es trivial, ya que pueden darse diferentes situaciones que compliquen esta labor, como variaciones en el formato del documento. Dos documentos pueden ser idénticos en contenido, pero estar en diferentes formatos (html, pdf, Word...). Una de las formas de detectar la similitud entre documentos consiste en convertirlos a **un mismo formato**, normalmente texto plano, utilizando alguna herramienta de conversión estándar. Posteriormente, cada documento se divide en una colección de partes o trozos formados por pequeñas unidades de texto (por ejemplo, líneas o frases). Después, a cada trozo se le aplica una función **hash** para obtener un identificador único. Si dos documentos comparten un número de trozos con igual identificador por encima de un umbral T , entonces se consideran documentos similares.

4.2.3. Estructuras de indexación clásicas

En los primeros sistemas, los índices se limitaban a contener un conjunto de **palabras clave** representativas del documento, pero actualmente el número de términos ha crecido demasiado. Como se ha dicho, en la indexación no se utilizan todos los términos (aunque hay excepciones), sino que se suele usar un subconjunto de términos. El **documento completo** se almacena aparte en repositorios o cachés, si es posible, aunque lo más habitual es que solo se almacene su ubicación (normalmente su URL, de *universal resource locator* en inglés).

La estructura más utilizada en la indexación de documentos es el **archivo invertido**, formada por dos componentes: el vocabulario y las ocurrencias. El **vocabulario** es el conjunto de todas las palabras diferentes del texto. Para cada una de las palabras del vocabulario, se crea una lista donde se almacenan las apariciones de cada palabra en un documento. El conjunto de todas estas listas se llama **ocurrencias** (Baeza-Yates y Ribeiro-Neto, 1999). Este mecanismo no es el único, sino que existen otros muchos, como los ficheros de firmas, basados en técnicas hash, árboles PAT y grafos.



Ejemplo

Fichero invertido después de que los documentos iniciales hayan sido preprocesados. Cada línea es un documento diferente:

Tabla 15

Documentos iniciales

Documento	Texto
1	"...Vivía cerca de León. Furioso con lo que le rodeaba, los hombres de la comarca lo odiaban..."
2	"...No tiene tanta furia el león como..."
3	"...León pertenece a Castilla..."
4	"...Los hombres y las comarcas castellanas..."

Fichero invertido para los documentos de la tabla a. Entre paréntesis se presentan las palabras antes de la lematización:

Tabla 16

Lematización o stemming

Número de índice	Término	Documento
1	vivía	1
2	león	1, 2, 3
3	furi (furioso, furia)	1, 2
4	rodeaba	1
5	hombres	1, 4
6	comarca (comarca, comarcas)	1, 4
7	odiaban	1
8	cast (castilla, castellanas)	3, 4

4.2.4. Consulta

El inicio de un proceso de búsqueda lo origina un problema que requiere información para poder resolverse. La carencia de esta información depende de la amplitud de conocimiento de cada usuario. Un usuario avezado en un tema concreto tendrá más claro que información solucionaría su problema y seguramente lo encontraría en un plazo de tiempo más corto. La aparición de un problema conlleva la demanda de información en el usuario para solucionarlo, y esta carencia de información origina lo que se denomina una **necesidad de información**.

Las personas buscan información basándose en su conocimiento previo, que es muy diferente de unas a otras. La necesidad de información puede ser definida como la representación implícita de un **problema** en la mente de los usuarios.

Se diferencia del problema, ya que cada usuario percibe las cosas de diferente forma, y ante un mismo problema varios usuarios pueden construir necesidades de información distintas.

Las necesidades de información se pueden clasificar en **necesidades verificativas**, sobre temas conscientes e imprecisas o mal definidas. La primera categoría se refiere a la situación en la que se buscan documentos con **propiedades conocidas**, por ejemplo, cuando se conoce el nombre del autor, el título, etc. En el segundo tipo **se conoce el tema** y es definible, pero con menor exactitud que en la primera categoría. En esta categoría una persona que busca información tiene algún nivel de comprensión de lo que busca. La tercera categoría son los casos en los que una persona desea encontrar **nuevo conocimiento** en dominios que no le resultan familiares.

Una necesidad de información se puede satisfacer de distintas formas. Es decir, el concepto de necesidad de información tiene una naturaleza ambigua. Debido a esta característica, se han comentado distintos problemas cuyo motivo es la **inexactitud de la necesidad de información**, como los problemas ASK (*anomalous state of knowledge* en inglés), ISK (*incomplete state of knowledge*) y USK (*uncertain state of knowledge*).

Cuando se aborda el desarrollo de un sistema de recuperación de información, se asume la idea de que las necesidades de información pueden describirse. La persona que quiere recuperar la información tiene que ser capaz de expresar la necesidad de información que demanda en forma de una **petición o consulta** (*query*). La petición es una representación de la necesidad de información del usuario en un lenguaje humano, casi siempre en lenguaje natural (no estructurado, como se ha comentado anteriormente en la recuperación de datos).

Sin embargo, esta consulta debe ser también **comprendible y procesable para el sistema de recuperación de información**. Evidentemente, la representación mental de la información que el usuario necesita para resolver su problema difiere enormemente de la información que recibe el SRI del usuario. Este proceso implica una adaptación de lo que el usuario cree que resolverá su problema a una expresión que represente lo que el usuario necesita encontrar.

Pero no basta con seguir este proceso para obtener la información que resuelva el problema. Si los resultados no satisfacen al usuario, puede ser necesario **repetir este proceso de forma cíclica**. Durante cada ciclo el sistema recibe realimentación del usuario con nueva información, formalizada en forma de nuevas consultas. En este proceso, se pueden distinguir a grandes rasgos 4 fases:

1. Fase de **exploración**. El usuario reúne la información que pueda serle útil en el proceso de búsqueda.
2. Fase de **construcción**. Se aprovecha la información adquirida en la fase anterior para reformular una nueva consulta.
3. Fase de **realimentación**. Si los resultados de la consulta formulada en la fase 2 no son satisfactorios, es necesario volver a realizar las fases 1 y 2 para refinar el resultado.
4. Fase de **presentación**. Se limita a la forma de representar los resultados.

Las técnicas de computación suave en la recuperación de información

A la hora de clasificar las diferentes líneas de investigación relacionadas con la recuperación de información y, más concretamente, con las posibilidades de las técnicas de computación suave en la recuperación de información en internet, se pueden utilizar diferentes criterios.

Podemos **clasificar en función de la parte del proceso de recuperación de información** en la que están centradas, en cuyo caso podemos distinguir los siguientes grupos de enfoques:

- Modelos de representación lógica de documentos.
- Lenguajes de especificación de consultas.
- Sistemas de evaluación de consultas.
- Sistemas de presentación y clasificación de resultados.
- Sistemas de retroalimentación de consultas.

Otra opción es distinguir los **enfoques según las técnicas utilizadas**:

- Utilización de ontologías.
- Estudio de asociaciones y relaciones entre términos.
- Construcción y utilización de perfiles de usuarios.
- Utilización de algoritmos de agrupamiento y clasificación.

Atendiendo al **objetivo y el ámbito de aplicación** de los sistemas de recuperación de información, podemos distinguir diferentes tipos de sistemas:

- Sistemas de búsqueda basados en consultas.
- Sistemas basados en directorios dinámicos.
- Sistemas de preguntas-respuestas (question-answering systems en inglés).
- Búsquedas conceptuales.

Todas estas categorías podrían ser a su vez subdivididas en función de si se han usado tecnologías típicas de computación suave, como la lógica borrosa, y técnicas de inteligencia artificial. Las posibilidades de estas técnicas en el campo de la recuperación de información están claramente constatadas en numerosos estudios.

Por último, cabe señalar que el propio profesor Zadeh, en los seminarios de BISC (Berkeley Initiative in Soft-Computing en inglés) siempre resaltaba la necesidad y actualidad de la línea propuesta. Para él, los motores de búsqueda existentes (con Google en la cúspide) tienen muchas capacidades notables; pero la que **no está entre ellas es la capacidad de deducción**, de sintetizar una respuesta a una consulta desde diferentes repositorios de información.

En los últimos años, se han realizado progresos impresionantes en la mejora del rendimiento de los motores de búsqueda mediante el uso de métodos basados en la lógica bivalente y en la teoría de la probabilidad basada en la lógica bivalente. Pero ¿se pueden utilizar estos métodos para añadir una capacidad de deducción no trivial a los motores de búsqueda, es decir, para actualizar los motores de búsqueda a sistemas de respuesta a preguntas? Una opinión es que la respuesta es "No". El problema tiene sus raíces en la naturaleza del conocimiento mundial, el tipo de conocimiento que los seres humanos adquieren a través de la experiencia y la educación.

Se reconoce ampliamente que el conocimiento del mundo desempeña un papel esencial en la evaluación de la pertinencia, la síntesis, la búsqueda y la deducción. Pero un tema básico que no se aborda es que **gran parte del conocimiento mundial está basado en la percepción**, por ejemplo, "Es difícil encontrar estacionamiento en París", "La mayoría de los profesores no son ricos" y "Es poco probable que llueva en pleno verano en San Francisco". El problema es que la información basada en la percepción es intrínsecamente borrosa y que la lógica bivalente es intrínsecamente inadecuada para tratar con la confusión y la verdad parcial.

Para enfrentarse a la imprecisión del conocimiento del mundo, se necesitan nuevas herramientas, como puede ser el **lenguaje natural preciso**. Este se basa en una lógica difusa y tiene la capacidad de tratar con la parcialidad de la certeza, la parcialidad de la posibilidad y la parcialidad de la verdad. Estas son las capacidades que se necesitan para poder aprovechar el conocimiento mundial para la evaluación de la pertinencia y para la síntesis, la búsqueda y la deducción.

A continuación, se describe qué papel puede desempeñar la **lógica borrosa** como técnica de computación suave para mejorar la búsqueda en este tipo de herramientas. La lógica borrosa puede proporcionar herramientas para la extracción y el uso de conocimiento procedente de tesauros y ontologías y permite formalizar sentencias, implementar capacidades de deducción en sistemas de tipo pregunta y respuesta, combinar valores borrosos y diferentes lógicas, mejorar algoritmos de agrupamiento y manejar las diferentes arquitecturas de un metabuscador.

Los principales **buscadores de la Red** basan sus criterios de búsqueda en aspectos léxicos y, en algunos casos, semánticos con respecto a los términos de la consulta. Debido a esto, son muchas las formas por las que se han tratado de mejorar los resultados de las búsquedas en línea, y es aquí donde las técnicas de computación suave pueden tener un papel importante. Prueba de ello es que en los últimos años se han propuesto múltiples soluciones basándose en este tipo de técnicas: soluciones dirigidas a la construcción de sitios flexibles y adaptables (basados en patrones web, perfiles de usuario, patrones de acceso, patrones de comportamiento de usuarios...) que utilizan técnicas de **aprendizaje automático**, otras centradas en la organización por grupos de documentos recuperados (destacando la importancia del uso de los algoritmos de agrupamiento en lugar de los algoritmos que se basan en los grupos temáticos predefinidos) o en otro tipo de aproximaciones que incluyen sistemas basados en lenguajes de consulta flexibles, o sistemas basados en reglas de asociación borrosa que ayudan al usuario a encontrar nuevos términos que utilizar en su consulta.

Por otro lado, existen distintos tipos de aproximaciones que se centran en la representación de documentos, para lo que se utilizan, en la mayoría de los casos, extensiones del **modelo de espacio vectorial** estándar. También es frecuente encontrar sistemas que utilizan las interrelaciones entre términos almacenadas en tesauros y ontologías, como *WordNet*, para construir redes semánticas de grupos de palabras.

Los **metabuscadores** surgen como una herramienta muy prometedora cuyo objetivo es el de mejorar los resultados de la búsqueda web. Para ello utilizan varios de los mejores motores de búsqueda, como Google o Yahoo, para después hacer una selección de los mejores resultados dados por dichas fuentes. Todos estos tipos de sistemas son muy diferentes a los que proponía Zadeh, que, como hemos visto, se centraban en el desarrollo de sistemas de pregunta y respuesta, un punto de vista muy interesante en problemas de recuperación de información.

Actualmente no hay muchos **buscadores comerciales con propiedades borrosas**. Tanto las técnicas de computación suave en general como de la lógica borrosa en particular pueden desempeñar un papel importante en los problemas relacionados con la búsqueda basada en tecnologías web. A continuación, se describen varios aspectos que podrían ser mejorados a través de las técnicas de computación suave.

Si hacemos la consulta “buscador borroso” en Google, serán muchos los resultados que aparezcan, pero, seleccionando aquellos que consideremos más relevantes, podemos observar que la idea de borrosidad que algunos buscadores comerciales implementan radica únicamente en el uso de funcionalidades de **emparejamiento sintáctico con propiedades borrosas**, es decir, tratan de corregir las palabras posiblemente mal escritas por el usuario mediante el uso de algún diccionario específico que contenga el buscador o al cual pueda tener acceso, y en el cual aparezcan las palabras que busca el usuario escritas de forma correcta. Como resultado, el buscador mandará una señal de aviso (texto escrito) que diría algo de la forma: “¿Quiso usted decir...?”. Aunque realmente esta funcionalidad es borrosa, es demasiado pretencioso hablar de buscador borroso simplemente por este detalle. Más concretamente, en buscadores famosos, el operador borroso de búsqueda permite al usuario escribir un trozo de la palabra que busca cuando no sabe a ciencia cierta cómo se deletrea la palabra completa.

Por ejemplo, muchos de los términos médicos o farmacéuticos son difíciles de deletrear. Quizá sabemos cómo suena la palabra, pero realmente no sabemos cómo se escribe. Tanto la búsqueda borrosa como los operadores *wild card* permiten encarar este problema. Por ejemplo, el buscador Netscape Search implementa funcionalidades borrosas. Así, si un usuario buscara el famoso antibiótico amoxicilina, podría hacer uso del operador borroso de búsqueda. Escribiría el trozo de palabra que supiera deletrear con certeza acompañado del símbolo “~” y el resto final de la palabra que ya no sabe si está deletreando bien o mal. Así, si hicieramos la consulta “amoxi~lilina”, el buscador devolvería satisfactoriamente aquellos documentos que contuvieran coincidencias del término buscado. Es similar a la búsqueda basada en caracteres *wild card*. Es posible combinar búsquedas en las que intervengan diferentes variantes de una palabra o deletreos incorrectos. Por ejemplo, muchos buscadores permiten tres caracteres *wild card*: el símbolo del dólar (\$), la interrogación (?) y el asterisco (*).

El papel de la lógica borrosa en los metabuscadores

Los **metabuscadores específicos** son hoy en día quizá una de las aplicaciones más usadas en lo que tiene que ver con el **acceso y análisis de datos**. Si observamos la publicidad en televisión nos damos cuenta de que gran parte de los anuncios son de metabuscadores específicos: elección de seguros, viajes, vuelos, hoteles... Los metabuscadores de carácter general no son tan abundantes debido a las dificultades que entraña su desarrollo y explotación.

Un metabuscador es un motor de búsqueda que, a diferencia de los buscadores conocidos, **no tiene una base de datos propia** donde indexar documentos.

He creado retornos de párrafo para que no parta este párrafo

Otra desventaja con respecto a los motores de búsqueda es que son más lentos debido a que, además de no tener base de datos propia, tienen que realizar un proceso de selección y elaboración de la lista de resultados bastante complejo. Sin embargo, disponen de un interfaz que permite al usuario consultar a la vez en diferentes motores de búsqueda, es decir, el metabuscador se encarga de recibir la petición del usuario, enviarla a **diferentes buscadores** y mostrarle después los resultados.

El principal problema que tienen los metabuscadores consiste en **combinar las listas de resultados** devueltos por los motores de búsqueda que utiliza, ya que en este proceso se deben clasificar y ordenar los documentos según su relevancia. Sin embargo, este tipo de sistemas tienen ciertas mejoras con respecto a los buscadores tradicionales. Solucionan algunos de los problemas que estos tienen, como el del *recall*, aunque hay otro tipo de inconvenientes que no han conseguido resolver, como el de la mejora de la **precisión**. Para llegar a conseguir mayor precisión en la búsqueda, se puede utilizar cualquiera de estos cuatro mecanismos: basados en el contenido, colaborativos, de conocimiento del dominio y basados en el uso de ontologías.

Los **métodos basados en el contenido** tratan de obtener una representación tan precisa como sea posible de las preferencias del usuario para después hacer una mejor evaluación y ordenación de las páginas devueltas. El **método colaborativo** se basa en establecer la similitud que puede haber entre usuarios para determinar la relevancia de la información. El **método basado en el conocimiento del dominio** se caracteriza por utilizar dos fuentes de información para proporcionar una mayor relevancia en los resultados: la ayuda del usuario y el conocimiento del dominio de búsqueda. Finalmente, el **método basado en ontologías** establece una jerarquía entre conceptos que permite concretar y mejorar la búsqueda.

Las **principales ventajas** que encontramos en los metabuscadores son las siguientes:

- Facilitan la consulta simultánea en múltiples buscadores. Es decir, permiten, por medio de una única consulta, obtener la lista ordenada de los documentos más relevantes que han devuelto los diferentes buscadores, evitándole al usuario la tarea de realizar la misma consulta en cada uno de ellos.
- Mejoran la eficiencia de recuperación. Dada la existencia de buscadores especializados en ciertos dominios, los metabuscadores pueden utilizarlos para mejorar los resultados finales y evitar así la información irrelevante que se pueda obtener de otros buscadores más generales.
- Resuelven el problema de la escalabilidad de la búsqueda en la Web.
- Incrementan la cobertura de búsqueda en la Web. Debido a la gran cantidad de documentos que hay en Internet, es imposible que un solo motor de búsqueda indexe la totalidad de links de la Web. Por lo tanto, con la combinación de diferentes buscadores, es posible cubrir un mayor número de documentos en las búsquedas.

Así mismo, un metabuscador también presenta las siguientes **ventajas potenciales**:

- Arquitectura modular. Las tecnologías utilizadas en un metabuscador se pueden dividir en módulos más pequeños y especializados que puedan ser paralelizados y ejecutados colaborativamente.
- Consistencia. Los buscadores actuales a menudo responden de manera muy diferente a la misma consulta según pasa el tiempo. Sin embargo, el metabuscador, al utilizar diferentes fuentes, tendrá menos variabilidad en los resultados, ya que se verá favorecido por aquellos buscadores que proporcionen resultados más estables.

- Mejora del factor recall. Habiendo obtenido los resultados de múltiples buscadores, se puede mejorar el número de documentos relevantes recuperados (el factor recall) con respecto al número total de documentos existentes.
- Mejora de la precisión. Diferentes algoritmos recuperan más documentos relevantes iguales, pero diferentes documentos irrelevantes. Basándose en este fenómeno, en caso de ser cierta esta teoría, cualquier algoritmo que dé prioridad a los documentos que aparecen en las primeras posiciones en los resultados de diferentes buscadores obtendrán una mejora en la recuperación. Este fenómeno se conoce como efecto coro.

A pesar de contar con toda esta serie de ventajas, los metabuscadores también tienen una serie de **inconvenientes**:

- La selección de la base de datos. Este problema se asocia con la selección del buscador más adecuado para recibir la consulta introducida, pues seleccionar el o los buscadores que devolverán buenos resultados para una consulta concreta no es nada sencillo. Por ejemplo, no tiene sentido la consulta "guitarra eléctrica" en un buscador especializado en literatura científica. Para intentar resolver esta desventaja, se propone el uso de medidas que indiquen la utilidad de cada base de datos con respecto a una consulta dada. Estos mecanismos se clasifican en tres categorías: métodos de representación amplia, métodos de representación estadística y métodos basados en aprendizaje.
- La selección de documentos. Una vez seleccionado el origen de los documentos, el problema consiste en determinar el número apropiado de documentos que es necesario devolver. Si se consideran demasiados, el coste computacional para determinar los mejores documentos y el coste de comunicación para obtenerlos puede ser excesivo. Se pueden establecer una serie de mecanismos para tratar de resolver este problema: decisión del usuario, peso (se obtienen mayor número de documentos del buscador que se considere el mejor), métodos basados en el aprendizaje (se basa en experiencias pasadas para determinar el número de documentos de cada buscador) y garantía de recuperación (trata de garantizar la recuperación de todos los documentos potencialmente útiles).
- Combinación de los resultados. El problema consiste en combinar los resultados de diferentes buscadores, teniendo en cuenta sus características y formas de evaluación, en una lista ordenada por relevancia. Además, existe la posibilidad de encontrar documentos devueltos que estén repetidos en diferentes buscadores. Las técnicas utilizadas para resolver este problema se centran en un ajuste de semejanza local (se basa en las características del buscador o la semejanza devuelta) y la estimación de una semejanza global (se evalúa o estima la semejanza de cada documento recuperado con la consulta original).

La traducción de una consulta a cada uno de los **lenguajes específicos de los buscadores** puede ser un factor importante en un metabuscador, dado que cada motor de búsqueda tiene su propio lenguaje de consulta.

Así pues, adaptar cada consulta al lenguaje de cada buscador parece necesario.

Hay muchas **arquitecturas de metabuscadores**. Habitualmente, esta estructura se descompone en una serie de **módulos más o menos específicos**:

- **Interfaz de usuario.** Es la encargada de recoger la consulta del usuario y posteriormente mostrar los resultados de la búsqueda. En algunos casos, la interfaz también contiene un sistema de refinamiento de la consulta, basado en el uso de algunas estructuras de conocimiento.
- **Selector de la base de datos.** Selecciona los buscadores que darán mejores respuestas a la consulta introducida por el usuario. Trata de evitar el envío masivo de consultas a los buscadores más lentos y con un elevado coste computacional.
- **Selector de documentos.** El objetivo es obtener el mayor número de documentos relevantes, evitando recuperar los no relevantes. Si se recupera un número excesivo de documentos no relevantes, la eficiencia de la búsqueda se verá afectada negativamente.
- **Emisor de consultas.** Es el encargado de establecer la conexión con el buscador y enviarle la consulta (o consultas), así como de obtener los resultados. Se utiliza habitualmente HTTP (hypertext transfer protocol en inglés) por el hecho de utilizar los métodos GET y POST. Sin embargo, existen buscadores que facilitan una interfaz de programación (API) para enviar consultas y que utilizan diferentes protocolos (Google usaba el protocolo SOAP en su API).
- **Agrupador de resultados.** Su función principal es combinar los resultados de los diferentes buscadores en una lista. Es esencial el uso de algún criterio de evaluación para establecer un orden en la lista que finalmente se muestre al usuario.

Hoy día es normal encontrar referencias bibliográficas acerca de **metabuscadores de tercera generación** (o motores de búsqueda de nivel 3). Estos trabajan de la siguiente forma: a petición del usuario, se crea una base de datos de nivel 3 a partir de los resultados obtenidos por los metabuscadores de nivel 2. Los metabuscadores (que representan el nivel 2 de los motores de búsqueda) utilizan motores de búsqueda estándar (de nivel 1) para encontrar los correspondientes resultados. Después se realiza un análisis de relevancia retroalimentado con estos resultados. Con esta colección de direcciones y documentos de texto, se desarrolla una base de datos (de nivel 3) que contiene solo documentos que sean relevantes en ese dominio de búsqueda para el usuario. En otras palabras, se crea una base de datos centrada específicamente en los campos de interés del usuario a partir de la cual podrá conseguir la información que necesita obteniendo unos resultados de gran calidad. Así pues, se plantea la necesidad de incluir perfiles de usuario y personalizaciones en futuros metabuscadores.

La idea es que un buscador sea **borroso** en el momento en que implemente búsquedas mediante **aproximaciones semánticas**, es decir, cuando incluya criterios de aproximación semántica a las consultas, no solo sintácticas. A continuación se presentan algunos aspectos a considerar.

El uso de diccionarios de sinónimos y tesauros (ontologías): búsqueda conceptual

Cuando un usuario realiza una búsqueda usando una única palabra, la búsqueda puede ser completada haciendo uso de un **diccionario de sinónimos**. El diccionario permite realizar búsquedas no solo teniendo en cuenta la palabra original sino además sus sinónimos. Se pueden establecer grados de sinonimia que después serán tenidos en cuenta para calcular el grado de relevancia de los documentos recuperados como respuesta a la consulta realizada por el usuario.

El proceso de búsqueda puede además ser mejorado usando **tesauros y ontologías**. En la actualidad, existen muchas ontologías referentes a diferentes dominios, que pueden mejorar diversos aspectos en ciertas aplicaciones. Todas ellas han sido hechas a mano siguiendo diferentes metodologías, tales como Methontology. En la otra cara de la moneda está la construcción automática de ontologías, que representa uno de los principales focos de investigación en la actualidad, donde hasta el momento los resultados obtenidos no son demasiados satisfactorios.



Enlace de interés

Possiblemente el tesauro más usado en la actualidad es WordNet, basado en las relaciones semánticas entre diferentes palabras.

www.wordnet.com

Las principales relaciones que gestiona **WordNet** son las de **sinonimia, hipónimia, hiperónimia, holónimia, merónimia**, etc. Así, un conjunto de sinónimos de una palabra puede ser agrupado en grupos llamados *synsets* y, como consecuencia, una palabra polisémica podrá pertenecer a diferentes *synsets*. La relación de hipónimia (y la de hiperónimia) se da entre diferentes términos entre los que se puede establecer una jerarquía. Así, por ejemplo, si la palabra *perro* es un tipo de canino, entonces el término *canino* es un hiperónimo del término *perro*. Este tipo de relaciones aportan información importante que permite expandir la consulta inicial del usuario **incorporando información con carácter semántico** al proceso de búsqueda, así como un mecanismo para la identificación del significado adecuado de los términos involucrados en la consulta.

Este tipo de sistemas normalmente requiere un mecanismo especial de *matching*, como el algoritmo de *ontomatching*, que compara los conceptos asociados a las palabras, o los sistemas de búsqueda que permiten **desambiguar significados** analizando la probabilidad de que coocurran ciertos conceptos. Se puede tratar el problema de la desambiguación estudiando el contexto local de las palabras y comparándolas con el contexto habitual de los distintos significados de las palabras. Este sistema requiere tener almacenado en un repositorio el contexto habitual de las palabras. La desambiguación desde el punto de vista de la computación suave podría verse como que las palabras no son desambiguadas en un único sentido, sino más bien en un conjunto de sentidos con sus correspondientes grados de relevancia. Hay modelos en los que se introduce el concepto de sinonimia relativa para definir un modelo de vectores basados en conceptos. En estos modelos, un término puede ser representado por un vector conceptual. Estos sistemas requieren repositorios de conceptos.

FIS-CRM es, por ejemplo, un modelo introducido por el autor y sus colaboradores para **representar los conceptos contenidos en cualquier clase de documentos**. Puede ser considerado como una extensión del modelo vectorial (VSM). Su principal característica es que se alimenta de información proveniente de un diccionario sinónímico borroso y varias ontologías temáticas. El **diccionario sinónímico** almacena el grado de sinonimia entre cada pareja de sinónimos y las **ontologías** guardan el grado de generalidad entre una palabra y otras más generales que esta. La clave del éxito del modelo FIS-CRM (en su aplicación a varios metabuscadores) radica primeramente en la construcción de los vectores base de los documentos teniendo en cuenta el número de ocurrencias de los términos (los vectores VSM) y el posterior reajuste de los pesos de los vectores con el propósito de representar el peso de las ocurrencias de conceptos, haciendo uso de la información disponible en el diccionario de sinónimos y las ontologías temáticas.

El proceso de reajuste conlleva el hecho de repartir las ocurrencias de un concepto entre los distintos sinónimos que convergen al mismo y que dan un peso a las palabras que representan un concepto más general que el que ellas mismas representan.

Sentencias de búsqueda y capacidades deductivas

Si la búsqueda incluye **frases**, además del diccionario de sinónimos, los tesauros y ontologías, será necesario el uso de **conectivas borrosas** adecuadas, para poder discernir, por ejemplo, en una consulta de la forma "A y B", cuándo A y B tienen información común o cuándo son totalmente independientes. Algo similar ocurre con la relación "A o B". Otro aspecto deseable es que se mantenga el significado de las palabras que se tienen en mente mediante la relación de sinonimia, para elegir la mejor función de similitud.

Pero el problema puede ser aún mayor si trabajamos con **relaciones causales**. En primer lugar, es muy difícil detectar una relación causal escrita (en un texto o una consulta). Por ejemplo, el texto "Lluvioso y oscuro" podría ser entendido por una persona como: "Si el tiempo está lluvioso, el cielo se oscurece". ¿Cómo puede un buscador distinguir la conectiva "y" y la causal? Ahora mismo es prácticamente imposible incluso si hay información relativa al contexto. Segundo, es muy difícil encontrar la función de implicación más adecuada para representar la sentencia (hay una gran variedad de implicaciones borrosas). La detección y la gestión de relaciones causales podrían ser muy importantes para desarrollar sistemas de pregunta y respuesta.

Para detectar las relaciones causales que existen en una colección de documentos, un punto de comienzo podría ser la detección de **frases condicionales**, pero esto no es tarea fácil. Descartes nunca pensó que su frase "Pienso luego existo" daría lugar a tal cantidad de conjetas e interpretaciones años después. En realidad, lo que él quiso decir es "Primero pienso y luego soy persona" o "Como tengo la capacidad de pensar, soy una persona". Incluso en esta ocasión, donde la intención de Descartes parece clara cuando expresó su máxima, no es tan fácil de interpretar y representar la información expresada en lenguaje natural, especialmente cuando conlleva frases complejas y giros complicados.

Con el fin de encontrar frases condicionales, se han desarrollado en el marco de nuestro grupo de investigación algunos **sistemas para detectar estructuras y clasificar frases causales** (Sobrino, Puente y Olivas, 2014), lo cual permite localizar, en términos de componentes básicos (verbos, adverbios, giros lingüísticos, etc.), algunas formas causales. Para realizar el análisis gramatical, tenemos, por un lado, que es posible separar ciertas relaciones causales considerando su forma verbal y, por otro, que es posible separar las relaciones atendiendo a los adverbios que aparecen en las frases. Ambos análisis dan lugar a algunas reglas causales que pueden ser usadas para extraer conocimiento de forma automática. De igual manera, cualquier estructura puede ser dividida en dos subestructuras que corresponden al antecedente y al consecuente de la relación causal, y existe un parámetro que mide el grado de certeza, conjeta o conformidad de dicha relación causal. En otras palabras, no es lo mismo una frase de la forma "Si gano la lotería, me compraré un coche", en la cual no hay duda de que, si el antecedente es verdadero, el consecuente también lo será; que tener una frase que diga "Si hubiéramos comprado un boleto en Sacramento, podríamos haber ganado la lotería", la cual deja muchas más dudas y conjetas, y no se puede asegurar, por tanto, que el cumplimiento del antecedente garantice el cumplimiento del consecuente.

Pero esto sigue siendo un problema de **procesamiento de lenguaje natural** muy complejo. Hay otras aproximaciones muy interesantes para la representación de frases condicionales mediante implicaciones borrosas. Por otro lado, también hay una idea basada en procesamiento de lenguaje natural y protoformas que podría ser una prometedora línea de investigación, tal y como propuso el profesor Zadeh.

Combinación de valores borrosos



Un metabuscador tiene que llevar a cabo una combinación de lógicas (los algoritmos que cada buscador usa) con la intención de **combinar las similitudes locales en una similitud global** o un orden final. Sin embargo, las similitudes locales no están basadas en criterios borrosos. Así, el orden de las páginas relevantes no es aproximado. Normalmente, los metabuscadores realizan las búsquedas de acuerdo a la importancia que ellos conceden a la pregunta del usuario y, en función de ella, evalúan los resultados para incorporarla a la lista final de resultados devueltos. En este caso, se tienen en cuenta **criterios** de mercado, y no criterios lingüísticos. Podría resultar interesante aplicar este criterio, primero, como se indicó antes, para conseguir búsquedas semánticas borrosas. En segundo lugar, los metabuscadores podrían crear una lista final de páginas recuperadas conforme a la relevancia proporcionada con los grados de confianza asociados a la lista local obtenida, no solo con las palabras tomadas de la consulta, y además usando términos relacionados como sinónimos, las medidas de similitud usadas en el cálculo de las relaciones lingüísticas y los operadores booleanos borrosos utilizados en las búsquedas. Cada búsqueda podría responder a diferentes lógicas borrosas realizadas por los buscadores, las cuales el metabuscador se encargaría de combinar para establecer una lista final de resultados. El uso de los enlaces proporcionados por el metabuscador para dar el orden de los resultados devueltos podría ser útil como banco de pruebas para experimentar distintas combinaciones hipotéticas de lógicas borrosas.

Otro problema que puede aparecer es cuando es necesario **agregar varios valores borrosos provenientes de distintas fuentes**. Dos palabras (conceptos) pueden tener más de una relación lingüística (cada una con su valor borroso), tales como la hiperonomia o la sinonimia. Por ejemplo, *balompié* y *fútbol* son sinónimos, pero el primer término es más general que el último. Puede existir una relación causal entre ambas palabras (conceptos). Más aún, se podría tener en cuenta una relación borrosa basada en la situación física de los términos (misma frase, párrafo, capítulo). Así pues, es necesario unir todos los valores borrosos en uno solo para aplicarlos en tareas de representación y búsqueda. Cómo agregar estos valores borrosos es un problema que no tiene solución conocida. Actualmente se usan los operadores estándar OWA (o derivados de ellos, como LOWA y WOWA).

Resultados del proceso de agrupamiento borroso

La clasificación de documentos o la **categorización de textos** (como se conoce en el mundo de la recuperación de información) es el proceso por el que se asigna un documento a un conjunto predefinido de categorías basándose en el contenido del documento. Sin embargo, las categorías predefinidas en un repositorio real de documentos no son conocidas.

Los métodos de agrupamiento de textos pueden ser aplicados para estructurar los conjuntos de documentos resultantes. Así, el usuario puede interactuar en el proceso de agrupamiento de los documentos. En definitiva, el proceso de agrupamiento permite la estructuración de la colección de documentos en un número reducido de grupos, donde los documentos de cada grupo tienen el suficiente grado similitud entre ellos.

Podemos enumerar los principales elementos que influyen a la hora de organizar un repositorio:

- Dimensionalidad. El clasificador puede manejar espacios de elementos de miles de dimensiones, por lo que es necesaria la capacidad de gestionar espacios de datos escasos o métodos de reducción de dimensiones.
- Eficiencia. Los algoritmos de agrupamiento documental deben ser eficientes y escalables. Además, el método debería ser preciso en la clasificación de nuevos documentos entrantes.
- Entendibilidad. El método debe proveer una descripción comprensible de los clusters descubiertos.
- Actualización. El clasificador debe ser capaz de actualizarse con cada nuevo documento que es archivado en el repositorio.

Existen muchos **algoritmos** de agrupamiento basados en técnicas de computación suave, por ejemplo, el *fuzzy C-means*, los mapas autoorganizados basados en arquitecturas de redes neuronales, como los mapas de Kohonen, etc. Actualmente los algoritmos de agrupamiento suave y las interfaces de agrupamiento dinámico son muy utilizados en las tareas de clasificación de los metabuscadores.

Arquitectura de un metabuscador

Como se ha comentado ampliamente, existen ciertos aspectos que provocan una considerable reducción de la eficacia de los procesos de búsqueda. Además de las conocidas limitaciones provocadas por el uso de palabras clave, se suman la inexperiencia de los usuarios en el uso de los motores de búsqueda y sus herramientas adicionales.

Así aparecen los metabuscadores, como una alternativa para tratar de paliar la baja precisión lograda hasta ahora por los buscadores. Además del uso de metabuscadores, también se proponen otras alternativas para tratar de mejorar el grado de relevancia de los resultados obtenidos por los buscadores, tales como las **técnicas de expansión de la consulta** basadas en el uso de términos relacionados semánticamente con los términos de la consulta original del usuario.

Existen muchas estrategias para llevar a cabo la técnica de expansión de la consulta, cada una con sus características propias. Básicamente, estos mecanismos pueden ser clasificados en automáticos, manuales e interactivos. Las **técnicas automáticas** tratan de añadir nuevos términos de forma automática relacionados semánticamente con aquellos que son introducidos manualmente por el usuario, para conseguir que el sistema devuelva una colección de documentos más acorde a la idea del usuario. Esta aproximación tiene algunos inconvenientes principalmente léxicos, como la polisemia (distintos significados para una misma palabra).

Se han desarrollado distintas implementaciones para tratar de identificar el significado adecuado de los términos de la consulta de forma automática. Estos algoritmos reciben el nombre de **desambiguación lingüística** (*word sense disambiguation* en inglés). El método de expansión de la consulta de forma interactiva requiere de la colaboración del usuario. El sistema propone una serie de términos al usuario que pueden estar relacionados con su consulta para que este elija aquellos que más se aproximen a la idea de lo que busca. Generalmente este tipo de sistemas utilizan una estructura en árbol que ordena los términos desde los más generales hasta los más específicos. Este tipo de sistemas suelen ser lentos e incómodos para el usuario porque requieren de un alto nivel de participación por su parte al responder a todas las preguntas a las que le somete el sistema. Por otro lado, la expansión de la consulta en muchas ocasiones se realiza utilizando estructuras de conocimiento como WordNet, aunque pueden ser utilizados muchos otros tesauros y ontologías.

Como ya se ha dicho, se han diseñado muchas arquitecturas sobre metabuscadores en las cuales se pueden identificar varios componentes comunes. Un ejemplo pueden ser aquellos componentes encargados de lanzar las consultas a los diferentes motores de búsqueda o bien los componentes encargados de calcular el grado de relevancia de los documentos recuperados.

La mayoría de las arquitecturas propuestas están basadas en el uso de **agentes específicos**, cada uno de ellos con diferentes funciones asignadas e intercomunicados a través de la Red. Cada agente tiene designada una función específica, pero trabajan de forma colaborativa con otros agentes para conseguir una reducción de la complejidad del sistema. Para ello es necesario un lenguaje para la comunicación entre agentes, conocido por todos ellos y que permita la colaboración entre ellos. Este lenguaje común puede ser, por ejemplo, ACL (de *agent communication language* en inglés).

La lógica borrosa podría desempeñar un papel fundamental en esta arquitectura basada en agentes, principalmente en la tarea de unir la información procedente de diferentes fuentes (agentes) y gestionar los resultados de forma eficiente y satisfactoria.

4.3. Minería de opiniones

La **minería de opiniones** o análisis de sentimientos (*opinion mining* en inglés) es uno de los temas de investigación más recientes en el ámbito de las ciencias de la computación. Hoy en día es uno de los campos más importantes, difíciles y demandados por la repercusión que tiene tanto para las empresas como para la sociedad. En las investigaciones desarrolladas por el grupo de investigación del autor de este manual, se han propuesto diversas aplicaciones y métodos. El artículo "Sentiment analysis: A review and comparative analysis of web services" (Serrano, Olivas, Romero, Herrera, 2015) ofrece una descripción del estado actual de la investigación y las aplicaciones en análisis de sentimientos y opiniones, de la que se presenta un breve extracto a continuación.

Las técnicas de recuperación de información textual se centran en el procesamiento, la búsqueda o la extracción de información objetiva. Los hechos tienen un componente objetivo; sin embargo, hay otros elementos textuales que expresan **características subjetivas**. Estos elementos son principalmente opiniones, sentimientos, valoraciones, actitudes y emociones, que son el foco del análisis de sentimientos. Todos ellos están estrechamente relacionados, aunque presentan ligeras diferencias. Este hecho implica el nacimiento de muchas tareas relacionadas en este nuevo campo de investigación, como la minería de opiniones, el análisis de subjetividad, la detección de emociones o la detección del *spam* de opinión, entre otros.

El análisis de sentimientos ofrece muchas oportunidades para desarrollar **nuevas aplicaciones**, especialmente debido al gran crecimiento de las herramientas disponibles, por ejemplo, las recomendaciones sobre los temas propuestos en fuentes como blogs y redes sociales. El **sistema de recomendaciones** puede funcionar teniendo en cuenta aspectos tales como las opiniones positivas o negativas sobre esos temas o productos. Los sitios web de opiniones podrían recopilar información de diferentes fuentes con el fin de resumir o componer una opinión global sobre un candidato, producto, etc., lo que sustituiría los sistemas que requieren explícitamente opiniones o resúmenes.

Los **sistemas de pregunta y respuesta** representan otro campo en el que las opiniones desempeñan un papel importante. La detección de preguntas sobre opiniones y sus posibles respuestas, así como su tratamiento, son esenciales para generar buenas respuestas. La detección de información subjetiva es realmente importante en campos relacionados con la argumentación en los que las frases objetivas suelen ser más valiosas. Aun así, ciertamente, uno de los campos más importantes en los que el análisis de sentimientos tiene un mayor impacto es la industria. Pequeñas y grandes empresas, al igual que otras organizaciones, como los gobiernos, desean saber lo que la gente dice sobre sus marcas, productos o miembros.

Como se ha dicho, el análisis de sentimientos es un concepto que abarca muchas tareas, como la extracción de sentimientos, la clasificación de sentimientos, la clasificación de subjetividad, el resumen de opiniones o la detección de *spam* de opinión, entre otros. Para llevar a cabo cualquiera de estas actividades, el análisis de sentimientos tiene que lidiar con muchos **desafíos**.

El primero es la **definición de los elementos** que intervienen. Por lo tanto, es necesario definir claramente conceptos como opinión, subjetividad o emoción, pero esta tarea no es fácil. Por ejemplo, de una manera sencilla, la opinión de un usuario puede ser considerada como un sentimiento positivo o negativo acerca de una entidad o de un elemento de esa entidad. Por otro lado, la subjetividad no implica necesariamente un sentimiento, sino que permite expresar sentimientos o creencias y, específicamente, nuestros propios sentimientos, creencias y emociones.

Estas definiciones tienen que ser formalizadas mediante expresiones matemáticas que puedan ser calculadas y utilizadas como entradas para los sistemas. Por lo tanto, el éxito del análisis de sentimientos depende principalmente de la **capacidad de extraer la información** necesaria de esas definiciones a partir de los textos para llevar a cabo esas tareas.

Así, las técnicas ya comentadas de procesamiento del lenguaje natural (**minería de textos**) son imprescindibles para conseguir buenos resultados en función de la tarea a realizar. Este es otro de los principales retos de este campo de investigación, junto con todos los problemas relacionados con la adaptación de técnicas típicas para clasificar o resumir, así como la creación de nuevas técnicas y algoritmos especializados en opiniones.

A pesar de la complejidad y la dificultad de este problema, muchas empresas y universidades están desarrollando **nuevas herramientas e instrumentos y servicios web** que tratan varios de los temas mencionados. Estos servicios podrían incluirse, especialmente para la investigación, en otras aplicaciones o plataformas sin necesidad de ser experto en análisis de sentimientos.

Los términos *minería de opiniones*, *análisis de sentimientos* y *análisis de subjetividad* (*opinion mining*, *sentiment analysis* y *subjectivity analysis* en inglés, respectivamente) se usan frecuentemente como sinónimos. Sin embargo, sus orígenes no son exactamente los mismos y algunos autores consideran que cada concepto presenta connotaciones diferentes, al igual que otros estrechamente relacionados, como el análisis afectivo.

4.3.1. Principales conceptos

Una opinión podría definirse simplemente como un sentimiento, una visión, una actitud, una emoción o una valoración positiva o negativa acerca de algo (producto, persona, evento, organización o tema) con respecto a un usuario o grupo de usuarios.

Siguiendo esa definición, una opinión puede ser matemáticamente definida como una 5-tupla $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$ donde:

- e_j representa una entidad.
- a_{jk} es el aspecto o característica k-ésima de la entidad e_j .
- so_{ijkl} es el valor sentimental de la opinión del propietario (holder en inglés) h_i en el aspecto a_{jk} de la entidad e_j en el tiempo t_l . h_i es el que tiene la opinión y t_l es el momento en que la opinión fue expresada

Ese valor puede ser **positivo**, **negativo** o **neutro**, o incluso necesitar una clasificación más granular.

Las opiniones **se pueden clasificar en diferentes grupos**. Por ejemplo, pueden ser opiniones regulares y comparativas. La mayoría de los dictámenes son periódicos y pueden subdividirse en dictámenes directos o indirectos. Las opiniones directas expresan una idea sobre una entidad o un aspecto de una entidad, mientras que las opiniones indirectas expresan una opinión sobre una entidad o un aspecto de una entidad basada en los efectos en otras entidades. Por otra parte, las frases comparativas expresan la semejanza entre entidades que consideran aspectos o características comunes. Además, las opiniones pueden clasificarse en explícitas o implícitas, dependiendo de si expresan ideas subjetivas u objetivas.

Aparte del sentimiento y la opinión, hay dos conceptos importantes cercanos a ellos, la **subjetividad** y la **emoción**. Una **oración subjetiva** puede expresar algunos sentimientos, puntos de vista o creencias personales; sin embargo, no implica necesariamente ningún sentimiento. Así, la diferencia entre oraciones objetivas y subjetivas es que una oración objetiva expresa algunos hechos o información sobre el mundo, mientras que una frase subjetiva expresa algunos sentimientos, puntos de vista o creencias personales. Un ejemplo podría ser la frase "Creo que se han ido". Sin embargo, la subjetividad a veces implica hasta cierto punto sentimientos cuando se trata de afecto, juicio, apreciación, especulación, acuerdo, etc.

Por otro lado, una **emoción** puede ser vista como una expresión de nuestros propios sentimientos y pensamientos subjetivos. Las emociones están muy cerca de los sentimientos. De hecho, la forma de medir la fuerza de una opinión está ligada a la intensidad de ciertas emociones, como el amor, la alegría, la sorpresa, la ira, la tristeza o el miedo. Un ejemplo podría ser la frase "Amo este coche", en la que el orador expresa su amor objetivo por su coche.

También es necesario comentar el concepto de **estado de ánimo**, que podría considerarse como una mezcla de sentimientos, emociones, sentimientos, que mueven al autor de un determinado texto a escribir ese comentario, observación, crítica, etc.

4.3.2. Tareas



Surgen muchas **tareas vinculadas al análisis de sentimientos**. Algunas de ellas están estrechamente relacionadas y es difícil separarlas claramente porque comparten muchos aspectos. Las más importantes son:

- 1. La clasificación de los sentimientos.** También llamada *orientación de sentimientos*, *orientación de opinión*, *orientación semántica* o *polaridad de sentimientos*. Se basa en la idea de que un documento o texto puede expresar una opinión de un titular sobre una entidad y trata de medir el sentimiento de ese titular hacia la entidad. Por lo tanto, consiste básicamente en clasificar las opiniones en tres categorías principales: **positivo**, **negativo** o **neutral**. Parece una tarea simple, pero es una tarea realmente compleja, especialmente cuando las opiniones provienen de múltiples dominios o idiomas. Esta tarea está estrechamente relacionada con la predicción de la valoración de los sentimientos, que consiste en medir la intensidad de cada sentimiento. Por ejemplo, se pueden utilizar diferentes escalas para medir una opinión, el rango $[-1, 1]$ donde -1 indica el grado negativo máximo y 1 el grado positivo máximo, o una escala de cinco estrellas en la que el usuario puede seleccionar cero estrellas para expresar la negatividad máxima o cinco estrellas en caso contrario.
- 2. Clasificación de subjetividad.** Consiste principalmente en detectar si una frase dada es subjetiva o no. Una frase objetiva expresa información factual, mientras que una frase subjetiva puede expresar otro tipo de información personal, como **opiniones**, **evaluaciones**, **emociones** y **creencias**. Además, las frases subjetivas pueden expresar lo positivo o lo negativo, pero no todas lo hacen. Esta tarea puede ser vista como un paso previo a la clasificación de los sentimientos. Una buena clasificación de subjetividad puede asegurar una mejor clasificación de sentimientos. Incluso se considera como un proceso más difícil que el de distinguir entre sentimientos positivos, neutros o negativos.
- 3. Resumen de opiniones.** Se centra especialmente en la extracción de las características principales de una entidad compartida en uno o varios documentos y los sentimientos sobre ella. Por lo tanto, se pueden distinguir dos perspectivas en esta tarea: resumen de un solo documento o de varios documentos. La **integración de documentos individuales** consiste en analizar los hechos presentes en el documento analizado, por ejemplo, cambios en la orientación de los sentimientos a lo largo del documento o vínculos entre las diferentes entidades o características encontradas, y mostrando principalmente aquellos textos que mejor las describen.

>>>



Por otro lado, en los **resúmenes multidocumento**, una vez detectadas las características y entidades, el sistema tiene que agrupar y/o ordenar las diferentes frases que expresan sentimientos relacionados con esas entidades o rasgos. Al final puede ser presentado como un gráfico o un texto que muestre las principales características o entidades y cuantifique el sentimiento de alguna manera en cada uno de ellos, por ejemplo, agregando intensidades de sentimientos o contando el número de sentimientos positivos o negativos.

- 4. Recuperación de opiniones.** Intenta recuperar documentos que expresan una opinión sobre una consulta determinada. En este tipo de sistemas, se necesitan dos puntuaciones para cada documento, la puntuación de relevancia frente a la consulta y la puntuación de opinión sobre la consulta. Normalmente se utilizan ambos para clasificar los documentos.
- 5. Sarcasmo e ironía.** Se centra en detectar afirmaciones con contenido irónico y sarcástico. Este es una de las tareas más complicadas en este campo, especialmente, debido a la falta de acuerdo entre los investigadores sobre cómo la ironía o el sarcasmo pueden ser formalmente representadas y definidas.
- 6. Otros.** Además de las actividades mencionadas anteriormente, se pueden destacar otras tareas relacionadas con el análisis de sentimientos, como la **detección de género o autoría**, que intenta determinar el género o la persona que ha escrito un texto u opinión, o la **detección de spam de opinión**, que trata de detectar opiniones o reseñas que contienen contenidos no confiables publicados para distorsionar la opinión pública hacia las personas, empresas o productos.

4.3.3. Técnicas

Las técnicas se suelen agrupar desde el punto de vista de las diferentes aplicaciones o retos que se pueden encontrar en análisis de sentimientos (AS) o en los principales temas de AS. Alguna clasificación los agrupa bajo **cinco grupos principales**: AS a nivel de documento, AS a nivel de frase, AS basado en aspectos, análisis comparativo de sentimientos y adquisición de léxico de sentimientos. Otras se centran principalmente en la agregación de opiniones, el *spam* de opinión y el análisis de contradicciones, especialmente aplicado a servicios web, por ejemplo, microblogs o streaming de datos, entre otros.

Existen cuatro tipos de técnicas de análisis de sentimientos: aprendizaje automático, basado en diccionarios, estadística y semántica.

Enfoques de aprendizaje automático

Como ya hemos visto, se pueden agrupar en dos categorías principales: técnicas supervisadas y no supervisadas. El éxito de ambos se basa principalmente en la selección y extracción del conjunto apropiado de características utilizadas para detectar sentimientos.

En esta tarea, las técnicas de procesamiento del lenguaje natural desempeñan un papel muy importante porque algunas de las características más importantes utilizadas son, por ejemplo:

1. Los **términos** (palabras o *n*-gramas) y su frecuencia.
2. La **información de la categoría gramatical**. Los adjetivos desempeñan un papel importante pero los sustantivos pueden ser significativos.
3. Las **negaciones**, que pueden cambiar el significado de cualquier oración.
4. Las **dependencias sintácticas** (análisis de árbol), que pueden determinar el significado de una oración.

Con respecto a las **técnicas supervisadas**, las máquinas de soporte vectorial (SVM), los clasificadores Bayes naíf y la máxima entropía son algunas de ellas, mientras que las **técnicas semisupervisadas y no supervisadas** se proponen cuando no es posible tener un conjunto inicial de documentos u opiniones etiquetados para clasificar el resto de ítems.

Además, los enfoques híbridos, que combinan técnicas supervisadas y no supervisadas, o incluso técnicas semisupervisadas, pueden ser útiles para clasificar los sentimientos.

Enfoques basados en el léxico

Se basan principalmente en un léxico de sentimientos, es decir, **una colección de sentimientos conocidos y precompilados**, términos, frases e incluso modismos, desarrollados para los géneros tradicionales de comunicación, como el Opinion Finder.

Estructuras aún más complejas, como **ontologías o diccionarios**, que miden la orientación semántica de las palabras o frases, pueden ser usadas para este propósito. Aquí se pueden encontrar dos subclasi ficaciones: enfoques basados en diccionarios y en corpus.

Los **enfoques basados en diccionarios** se basan generalmente en el uso de un conjunto inicial de términos (semillas), que normalmente se recogen y anotan de forma manual. Este conjunto crece buscando los sinónimos y antónimos en un diccionario.

Un ejemplo de ese diccionario podría ser WordNet, que se utilizó para desarrollar un tesauro llamado SentiWordNet.

El principal inconveniente de este tipo de enfoques es la incapacidad de hacer frente a orientaciones específicas de dominio y contexto. Aun así, podría ser una solución interesante dependiendo del problema.

Las **técnicas basadas en corpus** surgen con el objetivo de proporcionar diccionarios relacionados con un dominio específico. Estos diccionarios se generan a partir de un conjunto de términos de opinión de semillas que crecen a través de la búsqueda de palabras relacionadas mediante el uso de técnicas estadísticas o semánticas.

En los métodos basados en estadística como el **análisis semántico latente**, simplemente se puede utilizar la frecuencia de aparición de las palabras dentro de una colección de documentos. Por otro lado, los métodos semánticos, como el uso de sinónimos y antónimos, o las relaciones de tesauro, como WordNet, también pueden representar una solución interesante.

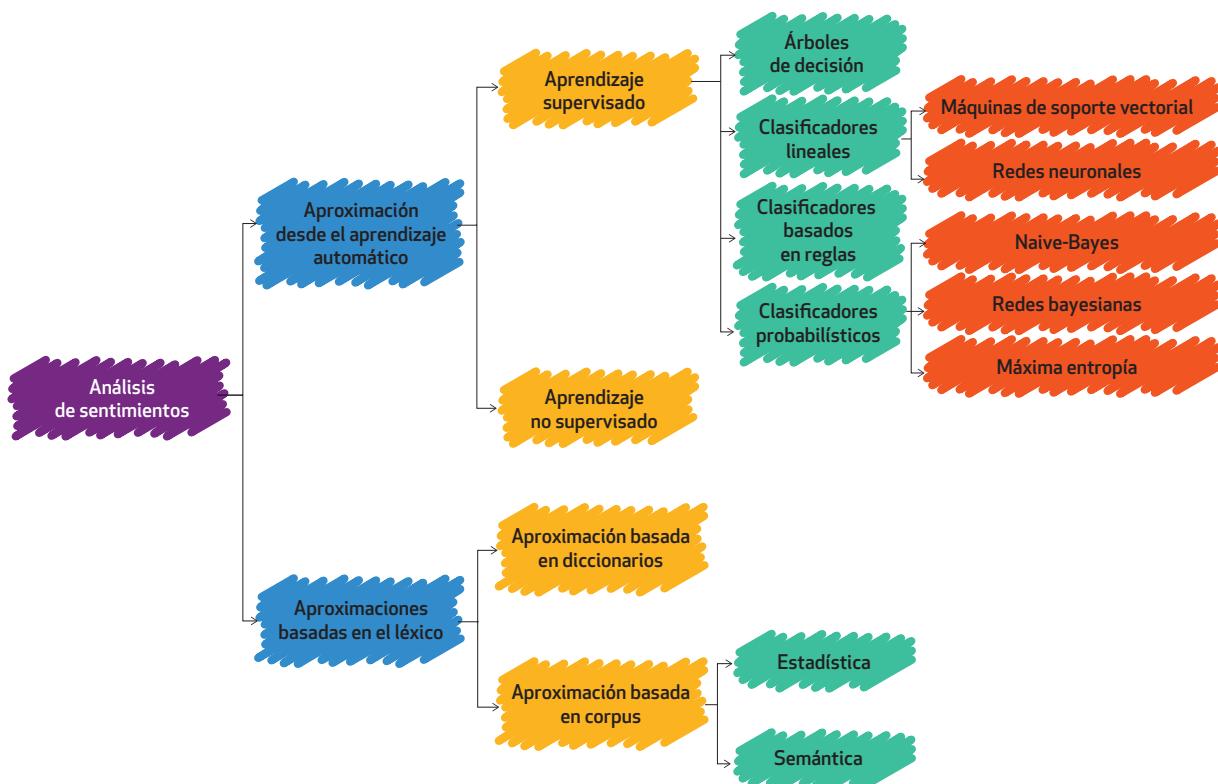


Figura 18. Técnicas usadas habitualmente para el análisis de sentimientos. Adaptado de "Sentiment analysis: A review and comparative analysis of web services", por J. Serrano, J. A. Olivas, F. P. Romero y E. Herrera, 2015, *Information Sciences*, 311, pp. 18-38.

Procesamiento del lenguaje natural y recuperación de información en el análisis de sentimientos

El análisis de sentimientos puede ser considerado como un problema de procesamiento de lenguaje natural (PLN) muy restringido, donde solo es necesario comprender los sentimientos positivos o negativos respecto a cada frase y/o las entidades o temas a tratar. Sin embargo, a pesar de ser un problema restringido, todos los trabajos en este campo, así como todos los trabajos en recuperación de información, siempre luchan contra los problemas no resueltos del PLN (manejo de la negación, reconocimiento de entidades con nombre, desambiguación del sentido de la palabra, etc.) que son esenciales para detectar recursos literarios como la ironía o el sarcasmo y, en consecuencia, para encontrar y valorar los sentimientos.

Uno de los aspectos principales del PLN son los diferentes niveles de análisis. Dependiendo de si el objeto de estudio es un texto o un documento completo, una o varias frases vinculadas, una o varias entidades o aspectos de esas entidades, se pueden realizar diferentes tareas de PLN y análisis de sentimientos.

Por lo tanto, es necesario distinguir **tres niveles de análisis** que determinarán claramente las diferentes tareas del análisis de sentimientos: nivel de documento, nivel de frase y nivel de entidad o aspecto.

El **nivel de documento** considera que un documento es una opinión sobre una entidad o aspecto de esta. Este nivel está asociado a la tarea de clasificación de sentimientos a nivel de documento. Sin embargo, si un documento presenta varias frases tratando diferentes aspectos o entidades, entonces el nivel de oración es más adecuado.

El **nivel de frase** está estrechamente relacionado con la tarea de clasificación de subjetividad, que distingue las frases que expresan información fáctica de las frases que expresan opiniones y puntos de vista subjetivos.

Finalmente, cuando se necesita información más precisa, entonces surge el **nivel de entidad o aspecto**. Es el nivel de grano más fino y considera un objetivo sobre el que el ponente de opinión expresa una opinión positiva o una opinión negativa. Este último nivel es posiblemente el más complejo, porque es necesario extraer con gran precisión muchas características tales como fechas o períodos de tiempo, las diferentes características, aspectos y entidades a tener en cuenta, así como las relaciones entre ellos, los opinadores y sus características.

Muchas propuestas siguen las mismas estrategias generales que otros trabajos de recuperación de información anteriores como los que hemos visto anteriormente, pero reemplazando varias variables estadísticas o semánticas por aspectos relacionados con los sentimientos. Por ejemplo, se propone el uso de la **cohesión léxica**, es decir, la distancia física entre las ubicaciones de los términos significativos o subjetivos para clasificar los documentos.

En otras propuestas se aplican métodos supervisados bien conocidos, como las redes neuronales o las máquinas de soporte vectorial (SVM) a la clasificación de sentimientos, las cuales han sido utilizadas profusamente en recuperación de información. En este caso también, la diferencia con otros trabajos sobre recuperación de información es la selección de características.



Aprendizaje automático

Es la rama de la inteligencia artificial en la que se diseñan mecanismos para dotar los sistemas computacionales de capacidad de aprendizaje, en el sentido de la capacidad de descubrir regularidades (patrones) en datos o situaciones anteriores y aplicarlos a nuevos problemas o situaciones análogas. Se pueden considerar diversos paradigmas y grupos de técnicas, como el aprendizaje supervisado (clasificación) y el no supervisado (*clustering*). En inglés se denomina *machine learning*.

Científico de datos

Es el profesional que posee conocimientos de computación, bases de datos, inteligencia artificial, aprendizaje automático, estadística, visualización, reconocimiento de patrones, sociología, psicología, KDD y minería de datos, para seleccionar y guiar las herramientas y técnicas más adecuadas para cada problema y los objetivos concretos en un proceso de análisis de datos. En inglés, esta figura se denomina *data scientist*.

Computación suave

La computación suave o blanda se diferencia de la computación convencional (dura) en que, a diferencia de ella, es tolerante a la imprecisión, la incertidumbre y la verdad parcial para lograr la trazabilidad, la robustez y el bajo coste de las soluciones. El modelo a seguir en este sentido es la mente humana. Los principales componentes de la computación suave son la lógica borrosa, la teoría de redes neuronales y el razonamiento probabilístico, con este último subsumiendo las redes de creencias, los algoritmos genéticos, la teoría del caos y partes del aprendizaje automático. En inglés se denomina *soft computing*.

Descubrimiento de conocimiento en bases de datos

Proceso (metodología) para, a partir de una base de datos (habitualmente estructurada), tratar de encontrar regularidades, 'patrones' en los datos que puedan ser representados formalmente y aplicados a situaciones futuras, con fines habitualmente de segmentación, predicción o pronóstico. En inglés se denomina *knowledge discovery in databases* (KDD).

Inteligencia artificial

La inteligencia artificial es la disciplina dentro del ámbito de la computación y los sistemas de información que pretende simular computacionalmente comportamientos humanos que pueden ser considerados como inteligentes. Se divide en varias ramas, como el aprendizaje automático, la ingeniería del conocimiento, la visión artificial o la robótica. En inglés se denomina *artificial intelligence*.

Lógica borrosa

La teoría de conjuntos borrosos fue introducida por Lotfi A. Zadeh. Bajo el concepto de conjunto borroso (*fuzzy set* en inglés) reside la idea de que los elementos clave en el pensamiento humano no son números, sino etiquetas lingüísticas. Estas etiquetas permiten que los objetos pasen de pertenecer de una clase a otra de forma suave y flexible. Uno de los objetivos de la lógica borrosa es proporcionar las bases del razonamiento aproximado que utiliza premisas imprecisas como instrumento para formular el conocimiento. También se llama *lógica difusa*.

Minería de datos

Se trata del análisis de datos en el que se parte de datos estructurados y casi siempre numéricos. El objetivo es encontrar regularidades (“patrones”) que permitan establecer modelos normalmente de predicción o de clasificación para situaciones futuras.

Minería de opiniones

Se trata del análisis de datos en el que se parte de colecciones de documentos de texto, habitualmente pequeños, como los típicos mensajes en redes sociales. El objetivo es manifestarse sobre la polaridad (bueno o malo) de un mensaje con respecto a un determinado tema para encontrar regularidades (patrones) semánticas (aunque normalmente solo se manejan desde el punto de vista lexicográfico) que permitan ayudar en tareas como la percepción de un producto o un político a través de las opiniones de los usuarios de las redes sociales. También se suele denominar *análisis de sentimientos*. Los equivalentes en inglés son (*opinion mining* o *sentiment analysis*).

Minería de textos

Análisis de datos en el que se parte de colecciones de documentos de texto. El objetivo es encontrar regularidades (“patrones”) semánticos (aunque normalmente solo se manejan desde el punto de vista lexicográfico) que permitan ayudar en tareas como el acceso, la búsqueda y la recuperación de información o la elaboración automática de resúmenes de dichos textos.

Sistemas basados en el conocimiento

La ingeniería del conocimiento es la parte de la inteligencia artificial encargada del desarrollo de sistemas basados en el conocimiento (SBC/KBS), como pueden ser los sistemas de ayuda a la decisión (DSS, de *decision support systems* en inglés). La tradición de los sistemas basados en el conocimiento (SBC) comenzó con los denominados *sistemas expertos*, sistemas computacionales que tratan de emular las capacidades de un experto en un tema basándose en la extracción del conocimiento del propio experto o grupo de expertos y transmitiéndoselo al sistema. Con la proliferación del almacenamiento y uso de datos de forma masiva, los SBC actuales suelen apoyarse en ambos pilares: expertos y datos.

Enlaces de interés



Wordnet

Possiblemente, el tesauro más usado en la actualidad es WordNet, el cual está basado en las relaciones semánticas entre diferentes palabras. Resulta imprescindible para hacer minería de texto y procesamiento de lenguaje natural.

<http://www.wordnet.com>

Zadeh (DEP), la lógica borrosa y el análisis de datos masivos

Recomendamos la lectura de este artículo (en el blog de la VIU) sobre la lógica borrosa.

<https://www.universidadviu.es/zadeh-d-e-p-la-logica-borrosa-analisis-datos-masivos>

Herramienta de lógica borrosa de Matlab

El siguiente enlace contiene una breve explicación de para qué sirve y dónde está encuadrada la herramienta de lógica borrosa de Matlab.

<https://es.mathworks.com/help/fuzzy/what-is-fuzzy-logic.html>

SciKit-Fuzzy

Scikit-Fuzzy es una colección de algoritmos de lógica borrosa para el uso con SciPy, escritos en Python.

<https://pythonhosted.org/scikit-fuzzy>

C Language Integrated Production System (CLIPS)

Esta es la página oficial de CLIPS, una herramienta gratuita creada en la NASA para el desarrollo de sistemas basados en reglas, muy fácil de usar y con una extensión denominada FUZZYCLIPS muy útil para desarrollar sistemas de reglas borrosos para implementar razonamiento aproximado.

<http://www.clipsrules.net>

Algunos 'mitos' sobre el aprovechamiento inteligente de los datos masivos

Recomendamos la lectura de este artículo (en el blog de la VIU) sobre algunos mitos del big data y la inteligencia artificial.

<https://www.universidadviu.es/mitos-aprovechamiento-inteligente-los-datos-masivos>

Bibliografía



Baeza-Yates, R., y Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Harlow: Addison-Wesley.

Black, M. (1937). Vagueness. An Exercise in Logical Analysis. *Philosophy of Science*, 4(4), 427-455.

Calatrava, C., Oruezábal, M. J., Olivas, J. Á., Romero, F. P., y Serrano, J. (2015). A Decision Support System for Risk Analysis and Diagnosis of Hereditary Cancer. En K. Atkinson y T. Sichelman (presidencia), *Proceedings of the 2015 International Conference on Artificial Intelligence*. Congreso llevado a cabo en San Diego.

Duda, R., Gaschnig, J., y Hart, P. (1981). Model design in the prospector consultant system for mineral exploration. En B. L. Webber y N. J. Nilsson (eds.), *Readings in Artificial Intelligence* (pp. 334-348).

Menger, K (1942). Statistical metrics. *Proceedings of the National Academy of Sciences of the United States of America*, 28(12), 535-537.

Olivas, J. Á. (2000). *Contribución al estudio experimental de la predicción basada en categorías deformables borrosas* (tesis doctoral, Universidad de Castilla-La Mancha).

Olivas, J. Á. (2002). La lógica borrosa y sus aplicaciones. *BOLE.TIC*, 24, 21-28.

Romero, R., Olivas, J. Á., Romero, F. P., Alonso, F., y Serrano, J. (2017). An Application of Fuzzy Prototypes to the Diagnosis and Treatment of Fuzzy Diseases. *International Journal of Intelligent Systems*, 32(2), 194-210.

Serrano, J., Olivas, J. Á., Romero, F. P., y Herrera, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311, 18-38.

Shortliffe, E. H., y Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3-4), 351-379.

Sobrino, A., Puente, C., y Olivas, J. Á. (2014). Extracting Answers from causal mechanisms in a medical document. *Neurocomputing*, 135, 53-60.

Zadeh, L. A. (1982). A note on prototype set theory and fuzzy sets. *Cognition*, 12, 291-297.



Autor
José Á. Olivas