"UN MÉTODO DE REPRESENTACIÓN DEL CONOCIMIENTO"

GARCÍA CAMARERO, E.

Históricamente aparecen las bases de datos cuando se ve la necesidad de que un mismo conjunto de datos sea utilizado por varios programas distintos, y por tanto ese conjunto de datos debe organizarse con independencia de los programas que los utilicen. En un principio esta organización se hacía con criterios apoyados en la estructura física de los soportes de información, después organizándose con criterios en los que aparecen aspectos semánticos de la información, así aparecen las estructuras de las bases jerárquicas, en red y relacionales. De esta forma las bases de datos se van convirtiendo en lo que deben ser: modelos del mundo exterior que corresponden a la representación del conocimiento del mundo sobre la que actuan los programas.

SENECA es una metodología de bases de datos que se inserta en la idea de conseguir una representación de conocimiento (relativo a un campo específico) atendiendo esencialmente a los aspectos semánticos de la información; para ello se requiere una descomposición del campo de estudio en sus partes esenciales y una integración posterior a través de relaciones explícitas que las unan. Junto a la información directamente contenida en el sistema puede obtenerse otra a partir de ella mediante inferencias lógicas o como resultados de las acciones que contenga.

En el presente artículo hacemos una introducción formal a SENECA, y damos una propuesta concreta a nivel experimental.

1. INTRODUCCION

En los sistemas informáticos se han distinguido siempre dos tipos de información: programas y datos. Planteado un problema se construía un programa para resolverlo y se daban los datos requeridos para encontrar la solución. El programa expresaba un algoritmo, y a él se prestaba casi toda la atención debido en general a su complejidad, y a la sencillez del conjunto de datos sobre los que actuaba. La organización de los datos estaba establecida en el interior del programa que los utilizaría. Cuando un mismo conjunto de datos debía de ser usado por varios programas distintos, aquellos debían organizarse de forma que pudieran ser utilizados por los distintos programas y los programas construirse teniendo en cuenta esta organización. En un principio, los criterios de organización de los datos se apoyaban en la forma de almacenamiento y en su localización, de forma que los programas pudieran encontrar y recuperar los datos necesarios para su ejecución.

Esta supeditación de la organización de los datos a la estructura física de almacenamiento dejó patente una rigidez de uso y puso de manifiesto que la estructuración de los datos no debía estar determinada por la estructura física de la memoria, por sus formas de acceso ni por su naturaleza formal, sino que debía realizarse teniendo muy en cuenta el contenido semántico de los datos para lograr una eficiente estructuración de los mismos. Así las bases de datos jerárquicos, en red y realcionales, son niveles de organización de datos en los que cada vez se tiene más en cuenta su componente semántico.

De esta forma las bases de datos se van convirtiendo en lo que deben ser: modelos del mundo exterior que corresponden a la representación del conocimiento del mundo sobre la que actuan los programas.

SENECA es una metodología que se inserta en esta idea, es decir en conseguir una representación del conocimiento atendiendo esencialmente a los aspectos semánticos de la información; considerando que para el análisis de esa información se requiere una descompo-

- E. García Camarero del Centre de Càlcul de la Universitat Complutense de Madrid. Ciudad Universitaria. Madrid.
- Article rebut el Febrer del 1980.

sición de la misma en sus partes esenciales y una integración posterior a través de las relaciones explícitas que las unen. Pero este análisis e integración debe realizarse teniendo en cuenta que toda la información debe estar disponible para una búsqueda utilizando criterios semánticos, y para construir información no dada explícitamente en la base mediante inferencias lógicas obtenidas a partir de la información existente.

La metodología SENECA la estamos desarrollando a tres niveles:

1) Nivel formal,
2) Nivel experimental,
3) Nivel de implementación.

El objetivo último sería la construcción de un sistema informático, en el que de forma continua se fuera incorporando información, y que esa información y la que de ella se pudiera inferir fuera facilmente accesible utilizando un lenguaje casi-natural. Estas inferencias se construirán de forma simple utilizando las propiedades de la propia red, o ejecutando las acciones que en la red irán incluidas. Así conseguiremos que el conjunto de datos no constituya una base inerte, sino propiamente una base activa de datos.

En la presente comunicación[1] haremos una sucinta exposición de las principales ideas y métodos usados en cada uno de los tres niveles a que hacíamos alusión más arriba, y del estado en que se encuentra nuestro trabajo.

2. NIVEL FORMAL

Hemos elegido como soporte formal para nuestra representación lo que se denomina redes semánticas, ya usadas con enfoques diversos por Simmons, Schank, Norman, etc... Nosotros damos la siguiente definición formal de red semántica, es decir la consideramos como un sistema formado por

$$G = (N, R, T_N, T_R, \phi_1, \phi_2)$$

consistente en un conjunto finito no vacío de nodos N, junto con una relación entre ellos RCNxN cuyos elementos denominaremos arcos, dos conjuntos de símbolos T_N, T_R denominados respectivamente etiquetas de los nodos y a-relaciones y dos funciones de asignación ϕ_1 y ϕ_2, tales que

$$\phi_1 : N \to T_N$$

$$\phi_2 : R \to T_R$$

Como ejemplo sencillo de esta formalización tendríamos, el siguiente grafo semántico

$$N = \{1,2,3,4,5,6,7\}$$

$$R = \{(1,2),(3,2),(3,4),(4,5),(5,2),(5,6),(7,2),(7,5),(7,6)\}$$

$$T_N = \{\square, o, \Delta\}$$

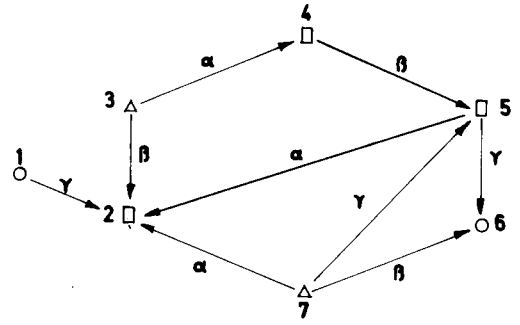
$$T_R = \{\alpha, \beta, \gamma, \delta\}$$

y las funciones ϕ_1 y ϕ_2 dadas por las siguientes tablas

N	1	2	3	4	5	6	7
$\phi_1(N)$	o	□	Δ	□	□	o	Δ

R	(1,2)	(3,2)	(3,4)	(4,5)	(5,2)	(5,6)	(7,2)	(7,5)	(7,6)
$\phi_2(R)$	γ	β	α	β	α	γ	α	γ	β

del que podríamos dar la siguiente representación gráfica:



Representación gráfica de un grafo semántico

Dos son los conceptos que queremos utilizar en un grafo semántico: la idea de proximidad y la idea de inferencia.

Para realizar la idea de proximidad introduciremos la noción de entorno y de entorno pautado o cualificado. Para definir entorno nos apoyaremos en las funciones de afluencia y confluencia, que designamos por ψ^+ y ψ^- respectivamente y que definimos así

$\psi^+ : N \to P(N)$

$\psi^- : N \to P(N)$

dandonos:

$\psi^+(n) = \{n_i | (n,n_i) \in R\}$

$\psi^-(n) = \{n_i | (n_i,n) \in R\}$

donde $n, n_i \in N$.

Teniendo en cuenta estas definiciones de ψ^+ y ψ^-, diremos que el entorno elemental de un nodo n, es el siguiente conjunto

$E(n) = \{n\} \cup \psi^+(n) \cup \psi^-(n)$

y el contorno de E(n) será

$C(n) = \psi^+(n) \cup \psi^-(n)$

y en general definiremos entorno y contorno de radio r, como sigue

$$E^r(n) = E^{r-1}(n) \cup [\underset{i}{\cup} E(n_i)] \qquad r \geq 2$$

donde

$i \in \{i | n_i \in C^{r-1}(n)\}$,

y

$$C^r(n) = E^r(n) - E^{r-1}(n) \qquad r \geq 2$$

Pero a veces esta idea de proximidad que introducimos con las nociones de entorno conviene pesarla cualitativamente, es decir, que consideraremos la proximidad alcanzada recorriendo determinadas a-relaciones, o secuencias de estas a las que llamamos pauta; una pauta es un elemento $x \in T_R^*$, es decir una sucesión arbitraria de a-relaciones. Así, llamaremos contorno pautado de pauta x, al siguiente conjunto

$$E_x^r(n) = E_x^{r-1}(n) \cup [\underset{i}{\cup} E(n_i)] \qquad r \geq 2$$

donde

$x'\beta = x$

con

$x' \in T_R^*$ y $\beta \in T_R$ y $n_i \in C^{r-1}(n)$; y consecuentemente llamaremos contorno pautado de pauta x al conjunto

$$C_x^r(n) = E_x^r(n) - E_x^{r-1}(n) \qquad f \geq 2$$

donde x' tiene el mismo significado que dimos antes.

Por otra parte la idea de inferencia la realizamos introduciendo tres tipos de reglas mediante las cuales dados dos arcos adyacentes en un determinado orden o posición respectiva, y pertenecientes a determinadas a-relaciones, obtenemos un nuevo arco con la indicación a la a-relación a la que pertenece. Estas reglas son de forma

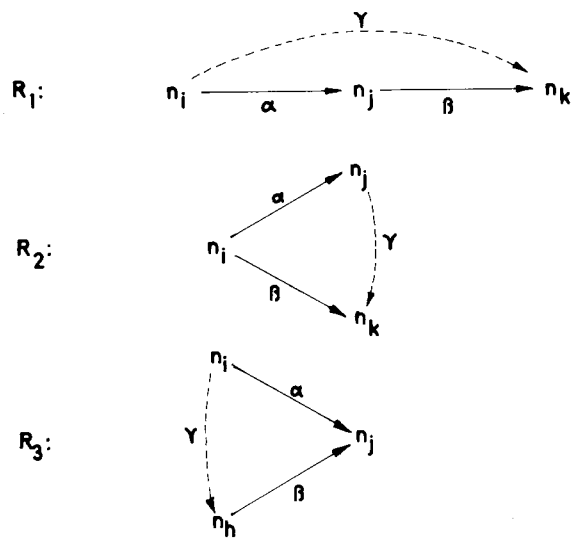
R.1 $(n_i,n_j) \in R_\alpha$, $(n_j,n_k) \in R_\beta \Rightarrow (n_i,n_k) \in R'_\gamma$

R.2 $(n_i,n_j) \in R_\alpha$, $(n_i,n_k) \in R_\beta \Rightarrow (n_j,n_k) \in R'_\gamma$

R.3 $(n_i,n_j) \in R_\alpha$, $(n_k,n_j) \in R_\beta \Rightarrow (n_i,n_k) \in R'_\gamma$

donde n_i, n_j, $n_k \in N$, α, β, $\gamma \in T_R$, y por R'_γ expresaríamos la γ-relación aumentada en caso necesario por el arco correspondiente que aparece en la expresión.

Gráficamente podríamos representar los anteriores tipos de la siguiente forma:



Representación gráfica de las reglas de inferencia por composición, afluencia y confluencia

Mediante esta regla podemos agregar o suprimir arcos a la red semántica sin que por ello se modifique la información potencial contenida en la misma, modificándose sólo la información explícitamente representada en --

ella. Así, por la aplicación de estas reglas podemos agregar arcos y construir nuevas redes semánticas equivalentes a la anterior y que diremos que es extendida o ampliada de ella, y podemos llegar a un punto en que no podamos agregar ningún arco mediante este proceso de inferencia obteniendo así una red que podemos llamar completa. De la misma forma podemos imaginar el proceso mediante el cual vamos suprimiendo arcos de un grafo que posteriormente podrían ser introducidos por inferencia; el nuevo grafo así obtenido lo llamaremos reducido del anterior. Se puede imaginar que por este proceso de reducción se llega a un grafo en el que no puede suprimirse ningún arco, sin perder información, al que llamaremos nucleo del grafo.

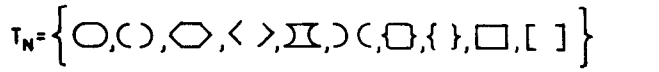
Por este proceso de inferencias podemos obtener información que no esté dada explícitamente por el grafo, o reducir el tamaño del grafo con las ventajas de implementación que ello conlleva.

3. NIVEL EXPERIMENTAL

Otro nivel de SENECA es el nivel experimental en el que definimos una red semántica específica con la que podemos experimentar en relación con elementos concretos.

La selección de los componentes de la red semántica se ha hecho por tentativas sucesivas orientadas a la modelización de un campo específico de la realidad. No hemos seguido procedimientos sistemáticos, aunque hemos procurado que los elementos elegidos sean de gran generalidad y por tanto aplicables a una gran variedad de campos.

Lo esencial en la elección del conjunto T_N particular, ha sido la partición de sus elementos en dos clases: una correspondiente a la representación de conceptos, y otra correspondiente a la representación de instancias de aquellos conceptos. Dentro de cada una de estas dos clases hemos supuesto que la parcela de conocimientos que queremos representar lo podemos dividir en los siguientes elementos: objetos, atributos, relaciones, clases, acciones. Para formar el conjunto T_N, hemos tomado las siguientes figuras:

T_N= 

donde las figuras cerradas representan elementos conceptuales, y las abiertas instancias de los mismos.

Para la terminación de las a-relaciones específicas que utilizaremos, hemos procurado determinar los tipos más elementales de relación que pueden vincular cada tipo de nodo; así hemos fijado, por ejemplo las siguientes a-relaciones:

A es una instancia de B
A es una subclase de B
A tiene como parte B
A es una propiedad que se aplica a B
A es la extensión de la relación B
A es un elemento de la clase B
A es argumento de una acción o de una relación B
A es el resultado de la acción B

abreviadamente escribimos

A es un B
A clas B
A tcp B
A ap B
A ext B
A el B
A arg B
A res B

de forma que el conjunto T_R, estaría formado por los siguientes elementos:

T_R = {esun,clas,tcp,ap,ext,el,arg,res}

Es claro, que no todo par de nodos puede estar vinculado por cualquier tipo de a-relación, y que el tipo de vinculación depende del tipo de nodo que vinculan. En la tabla I, se dan las a-relaciones que pueden vincular cada dos tipos de nodos, habiendo nodos que no pueden estar vinculados por ninguna a-relación.

En la figura 1 aparece la representación gráfica de un ejemplo sencillo de red semántica en la que se dejan patentes los dos planos esenciales elegidos en nuestra representación: el plano conceptual y el plano instancia.

Las reglas de inferencia que enunciamos de una forma general en el nivel conceptual, para el caso de la red experimental aquí pro-

Tabla I.
Tabla de las a-relaciones posibles entre ca-
ca par de tipos de nodos.

	⬭	( )	⬡	< >	⋈	) (	▭	{ }	▭	[ ]
⬭	tcp clas	•	•	•	arg	arg	elem	•	arg	•
( )	esun	tcp	•	•	•	arg	•	elem	•	arg
⬡	ap	ap	tcp ap clas	ap	ap	ap	elem ap ext	ap	ap arg	ap
< >	ap	ap	ap esun	tcp ap	ap	ap	ap	ext ap elem	ap	ap arg
⋈	•	•	•	•	tcp clas	•	ext	•	•	•
) (	•	•	•	•	esun	tcp	•	ext	•	•
▭	•	•	•	•	arg	arg	tcp clas elem	•	arg	•
{ }	•	•	•	•	•	arg	esun	tcp elem	•	arg
▭	res	•	res	•	arg	•	elem res	•	tcp clas	•
[ ]	•	res	•	res	•	arg	•	elem res	esun	tcp

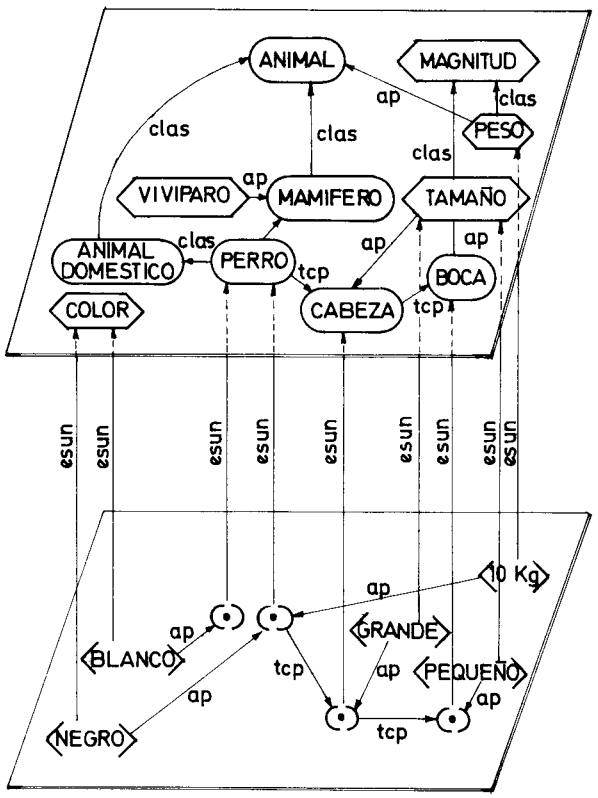


Fig. 1
Representación de la red semántica dejando - patente dos planos esenciales: el superior o conceptual, el inferior o concreto formado - por instancias.

puesta, se concretizan de la forma que expre samos a continuación:

1. La regla de composición *, queda definida de la siguiente forma

$$R_\alpha * R_\beta = \{(x,y) \mid (x,z) \in R_\alpha \Delta (z,y) \in R_\beta\}$$

donde

$$R_\alpha = \phi_2^{-1}(\alpha) \quad y \quad R_\beta = \phi_2^{-1}(\beta)$$

los nuevos pares (x,y) son agregados a la relación R_γ de la siguiente forma

$$\text{Si } \alpha = \beta \Rightarrow \begin{cases} \gamma = \alpha = \beta \\ R'_\gamma = R_\gamma \cup \{(x,y)\} \end{cases}$$

$$\text{Si } \alpha = \text{clas}, \ \beta = \text{tcp} \Rightarrow \begin{cases} \gamma = \text{tcp} \\ R'_\gamma = R_\gamma \cup \{(x,y)\} \end{cases}$$

$$\text{Si } \alpha = \text{esen}, \ \beta = \text{clas} \Rightarrow \begin{cases} \gamma = \text{esun} \\ R'_\gamma = R_\gamma \cup \{(x,y)\} \end{cases}$$

lo que daría representado gráficamente en la figura 2.

2. La regla de confluencia >, queda definida de la siguiente forma

$$R_\alpha > R_\beta = \{(x,y) \mid (x,z) \in R_\alpha \Delta (y,z) \in R_\beta\}$$

donde R_α, R_β tiene el mismo significado - que antes; los nuevos pares (x,y) que son agregados a la relación R_γ para formar R'_γ se obtienen de la siguiente forma:

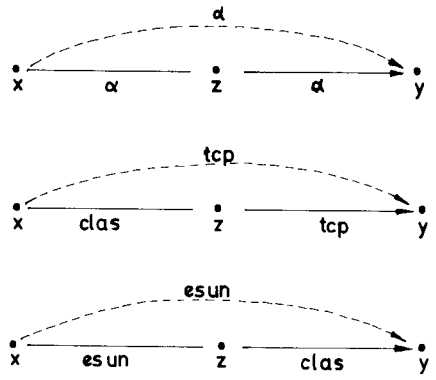


Fig. 2
Representación gráfica de las inferencias , por composición.

$$\text{Si } \alpha=ap,\ \beta=esun \Rightarrow \begin{cases} \gamma = ap \\ R'_\gamma = R_\gamma \cup \{(x,y)\} \end{cases}$$

$$\text{Si } \alpha \text{ ext}, \ \beta \text{ elem} \Rightarrow \begin{cases} \gamma = ap \\ R'_\gamma = R_\gamma \cup \{(x,y)\} \end{cases}$$

En la figura 3, damos una representación gráfica de esta situación.

3. La regla de afluencia $<$, la definimos de la siguiente forma

$$R_\alpha < R_\beta = \{(x,y) \mid (z,x) \in R_\alpha \wedge (z,y) \in R_\beta\}$$

la forma de agregar nuevos pares a la relación R_γ, se realiza así:

$$\text{Si } \alpha=clas,\ \beta=ap \Rightarrow \begin{cases} \gamma = ap \\ R'_\gamma = R_\gamma \cup \{(x,y)\} \end{cases}$$

Con estas reglas sabemos que arcos pueden ser inferidos o suprimidos, manteniendo el mismo contenido semántico potencial en la red.

4. NIVEL IMPLEMENTACION

En este nivel de implementación se han desarrollado en mayor o menor escala cuatro ejemplos. Uno relativo a la descripción de un corpus de textos literarios, en el que se han considerado tanto la descripción bibliográfica, la descripción morfológica de los textos, así como el entorno biográfico e histórico en el que se produjo el texto. Otro relativo a la descripción de un objeto de interés histórico así como el contexto histórico en el que se situa, el tercer ejemplo se realizó para constrastar la potencia de representación con respecto a las bases relacionales y se tomó un ejemplo de gestión dado por un libro clásico de bases de datos relacionales y el cuarto es otro ejemplo relativo a la gestión académica de estudiantes.

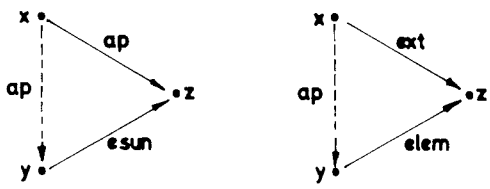


Fig. 3
Representación gráfica de las inferencias por concluencia.

Se ha realizado implementación de las redes correspondientes (cada una de ellas de mayor a menor amplitud pero todas experimentales) en lo referente a la representación en memoria. También se han estudiado algoritmos para manejar los entornos y entornos pautados, que son básicos a la hora de búsqueda de información, así como los algoritmos correspondientes a las inferencias específicas que permiten obtener información adicional u optimizar el empleo de la memoria.

Para la representación de las preguntas también se utilizan redes semánticas en las que figuran uno o varios nodos incógnitas y estas podrán estar afectadas de cuantificadores. También se ha desarrollado un procedimiento efectivo para la determinación de las respuestas, es decir de los nodos que satisfacen las preguntas, apoyándose en los nodos conocidos que figuran en la red pregunta.

5. BIBLIOGRAFIA

A continuación damos una lista de algunos trabajos en los que pueden encontrar detalles sobre las cuestiones tratadas sintéticamente en la presente comunicación.

* GARCIA CAMARERO, E., VERDEJO, M.F.: "Un sistema pregunta-respuesta en castellano sobre un corpus literario". Bol. Centro de Cálculo Universidad Complutense, n° 32, Madrid, Mayo 1978, p. 4-12.

* GARCIA CAMARERO, E., GARCIA SANZ, J., VERDEJO, M.F.: "Construcción de una base activa de datos lingüísticos. Primera Memoria". (Memoria interna, presentada a FUNDESCO), Madrid, Julio 1978, 177 pp.

* GARCIA CAMARERO, E., VERDEJO, M.F., VIRBEL, J.: "Una aplicación de SENECA a un dominio arqueológico". Bol. Centro Cálculo Universidad Complutense, n° 33, Madrid, Diciembre 1978, pp. 1-11.

* GARCIA CAMARERO, E., GARCIA SANZ, J., VERDEJO, M.F.: "Representación del conocimien

to mediante redes semánticas". Actas de la Convención Informática Latina CIL 79, Barcelona 1979, pp. 457-466.

* "Red semántica para la gestión de una base de datos lingüísticos". Actas de la Convención Informática Latina, CIL 79, Barcelona 1979, pp. 467-478.

* "SENECA: Semantic Networks for Conceptual Analysis". A aparecer en los Proceedings de la Conference on Data Bases in Humanities and Social Sciences, celebrada en Hannover, N.H. Agosto 1979.

* "Resolución de sistemas de ecuaciones booleanas con variables en el conjunto de los contornos de un grafo". Congreso de la AEIA, Madrid, Octubre 1979.

* "Un método de respuestas a preguntas formuladas a un grafo semántico". Congreso de la AEIA, Madrid, Octubre 1979.

* NORMAN, D., RUMELHART, J.: "A process model for long term memory". E. Tulving Donallson edit. 1972.

* SCHANK, K.F.: "Conceptual Information Processing". North Holland. Amsterdam. 1975.

* SIMMONS, R.F.: "Semantic Networks, their computation and use ofr understanding english sentences". en Schank & Colby, ed. Freeman. San Francisco. 1973.

6. NOTAS

1) Esta comunicación es una síntesis expositiva de los trabajos realizados por el autor y los Sres. Verdejo y García-Sanz del Centro de Cálculo de la Universidad Complutense de Madrid, y fué presentada al "Colloque sur la Représentation des Connaissances et du Raisonnement dans les Sciences de l'Homme et de la Societé" en Saint Maximin los días 17-19 de Septiembre de 1979, organizado por el Laboratoire d'Informatique pour les Sciences de l'Homme del C.N.R.S. de Francia.